ARO 14483.12-M

Report DRXCRO-PR P-14483-M

LEVEL II (12)

APPROXIMATE LINEAR REGULATOR AND KALMAN FILTER

AD A092176

Leang S. Shieh
Willon B. Wai
Department of Electrical Engineering
University of Houston
Houston, Texas 77004

September 1, 1980

Final Report for Period 1 June 1977 - 31 August 1980

Prepared for

Mathematics Division
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

Report DRXCRO-PR P-14483-M

APPROXIMATE LINEAR REGULATOR AND KALMAN FILTER

Leang S. Shieh
Willon B. Wai
Department of Electrical Engineering
University of Houston
Houston, Texas 77004

September 1, 1980

Final Report for Period 1 June 1977 - 31 August 1980

Prepared for

Mathematics Division
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

18 ARO 19 14483.12M

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: DRXCRO-PRP-14483-M
2. GOVT ACCESSION NO.: AD-A092176
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): APPROXIMATE LINEAR REGULATOR AND KALMAN FILTER
5. TYPE OF REPORT & PERIOD COVERED: Final Report, 1 June 1977 — 31 August 1980
6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Leang S. Shieh, Willon B. Wai
8. CONTRACT OR GRANT NUMBER(s): DAAG-29-77-G-0143, DAAG 29-79-C-0178, DAAG29-77-G-0143

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Department of Electrical Engineering
University of Houston
Houston, Texas 77004

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS:
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park NC 27709

12. REPORT DATE: 1 September 1980
13. NUMBER OF PAGES: 80

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
Mathematics Division
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

15. SECURITY CLASS. (of this report): unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report):
Approved for public release: distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):

18. SUPPLEMENTARY NOTES:
THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
Linear Regulator, Kalman Filter, Geometric-Series Method, Discrete-time Model, Transition Matrix.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number):
Practical dynamic systems constantly face unpredictable fluctuations and disturbances for which Kalman filter has been shown to be effective in estimating the states from the outputs corrupted by white noises. This is the Kalman filtering problem. On the other hand, the Linear regulator problem, which is the mathematical dual of the Kalman filtering problem, plays an important role in modern optimal control theory. Both problems can be formulated as quadratic synthesis problems.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

401227

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A gometric-series approach is used to approximate the exponentials of Hamiltonian matrices for the quadratic synthesis problems. The approximants of the discretized transition matrices are then used to construct piecewise-constant gains and piecewise time-varying gains for approximating time-varying optimal gains and time-varying Kalman gains. Simple and fast algorithms are developed and can be easily implemented on a low cost minicomputer or microprocessor.

The proposed methods have been successfully applied to the analysis of practical control systems.

Other new findings of this research are reported in the appendix.

ABSTRACT

Practical dynamic systems constantly face unpredictable fluctuations and disturbances for which the Kalman filter has been shown to be effective in estimating the states from the outputs corrupted by white noises. This is the Kalman filtering problem. On the other hand, the Linear regulator problem, which is the mathematical dual of the Kalman filtering problem, plays an important role in modern optimal control theory. Both problems can be formulated as quadratic synthesis problems.

A geometric-series approach is used to approximate the exponentials of Hamiltonian matrices for the quadratic synthesis problems. The approximants of the discretized transition matrices are then used to construct piecewise-constant gains and piecewise time-varying gains for approximating time-varying optimal gains and time-varying Kalman gains. Simple and fast algorithms are developed and can be easily implemented on a low cost minicomputer or microprocessor.

The proposed methods have been successfully applied to the analysis of practical control systems.

Other new findings of this research are reported in the appendix.

TABLE OF CONTENTS

Appendix (A) Publications

(1) L. S. Shieh, Y. J. Wei, H. Z. Chow and R. E. Yates, "Determination of Equivalent Dominant Poles and Zeros Using Industrial Specifications," IEEE Trans. on Industrial Electronics and Control Instrumentation, Vol. IECI-26, No. 3, pp. 125-133, August 1979.

(2) Y. J. Wei and L. S. Shieh, "Synthesis of Optimal Block Controllers for Multivariable Control Systems and Its Inverse Optimal-Control Problem", Proceedings of the IEE (England) Vol. 126, No. 5, pp. 449-456, May 1979.

(3) L. S. Shieh, M. Datta-Barua and R. E. Yates, "A Method for Modelling Transfer Functions Using Dominant Frequency-Response Data and Its Applications," International Journal of Systems Science, Vol. 10, No. 10, pp. 1097-1114, October 1979.

(4) L. S. Shieh, R. E. Yates, J. P. Leonard, and J. M. Navarro, "A Geometric-Series Approach to Modelling Discrete-Time State Equations from Continuous-Time State Equations," International Journal of Systems Science, Vol. 10, No. 13, pp. 1415-1426, 1979.

(5) L. S. Shieh and A. Tajvari, "Analysis and Synthesis of Matrix Transfer Functions Using the New Block-State Equations in Block-Tridiagonal Forms," IEE Proc. (London) Vol. 127, PtD. No. 1, pp. 19-31, January 1980.

(6) L. S. Shieh, M. Datta-Barua, R. E. Yates, and J. P. Leonard, "Computer-Aided Methods to Redesigning the Stabilized Pitch Control System of a Semi-Active Terminal Homing Missile," Accepted for publication in International Journal, Computers and Electrical Engineering, 1980.

(7) L. S. Shieh, W. B. Wai and R. E. Yates, "A Geometric Series Approach for Approximation of Transition Matrices in Quadratic Synthesis," Accepted for publication in ASME Journal of Dynamic Systems, Measurement, and Control, 1980.

Appendix (B) Presentations

(1) L. S. Shieh and A. Tajvari, "Some Properties and Applications of a New Matrix Sturm Series and a New Block Canonical Form of a Matrix Transfer Function", Presented at the 22nd Midwest Symposium on Circuits and Systems, June 1979.

(2) L. S. Shieh, W. B. Wai, and R. E. Yates, "Approximate Kalman Filters," Accepted for presentation at the 1980 JACC (Joint Automatic Control Conference), August 1980.

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Historical Review

The development of control theory has become one of the cornerstones in modern technology. Classical control system design is generally a trial-and-error process in which various methods of analysis such as Nyquist, Bode and Routh-Hurwitz criteria were used iteratively to determine the design parameters of a deterministic system. During the postwar development, control engineers were faced with several problems which required a very stringent performance. Many of the control processes they dealt with became extremely complex. For example, the design of spacecraft attitude with minimum fuel expenditure requirement is not applicable to the classical methods. Such a problem has led to a new formulation of an optimal control system. This system is as much a branch of applied mathematics as of control engineering. Methods of design require sophisticated mathematical tools such as differential equations, calculus of variation and dynamic programming. The objective of optimal control theory is to determine the control laws which will make a system satisfy its physical constraints and at the same time minimize the performance criteria. The practical applications of optimal control ideas in the various space missions make the dream of investigating the universe come true. In recent years, the rapid development of powerful minicomputers and microprocessors makes the industrial applications of optimal control systems popular, and they will undoubtedly become increasingly important in the future. The linear regulator problem and Kalman filter problem are reviewed as follows.

1.2 Linear Regulator Problem

An optimal control problem can be illustrated in the following fashion [1]:

Given a system equation,

$$\dot{x}(t) = f(x(t),u(t),t)\,, \tag{1-1a}$$

find an admissible control, $u^*(t)$, which causes the system to follow an admissible trajectory, $x^*(t)$, that minimizes the performance measure,

$$J = h(x(t_f),t_f) + \int_{t_0}^{t_f} g(x(t),u(t),t)dt\,. \tag{1-1b}$$

$u^*(t)$ is called an optimal control and $x^*(t)$ an optimal trajectory.

If the system is linear and time-varying, its state equation is:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)\,, \tag{1-2a}$$

where $x\epsilon R^n$, $A\epsilon R^{n\times n}$, $B\epsilon R^{n\times p}$ and $u\epsilon R^p$

The performance index becomes

$$J = \frac{1}{2}\, x^T(t_f)H\, x(t_f) + \frac{1}{2}\int_{t_0}^{t_f} [x^T(\lambda)Q(\lambda)x(\lambda)+u^T(\lambda)R(\lambda)u(\lambda)]d\lambda\,, \tag{1-2b}$$

where H and Q are real, symmetric, positive semi-definite matrices, and R is a real, symmetric, positive-definite matrix. The initial

time, t_0, and the final time, t_f, are specified, and u(t) and x(t) are not constrained by any boundaries. This is called a linear regulator problem. The speed control system of a turbine-generator set in a power station, and the level control system of plate glass manufacturing are examples of such problems since the generator speed and liquid level need to be as near a constant as possible. The Hamiltonian of the system (1-2) is:

$$H(x(t),u(t),J_x^*,t) = \frac{1}{2}\, x^T(t)Q(t)x(t) + \frac{1}{2}\, u^T(t)R(t)u(t) + J_x^{*T}(x(t),t)[A(t)x(t)+B(t)u(t)]. \qquad (1\text{-}3)$$

By use of the Hamilton-Jacobi-Bellman equation [2], a necessary condition for u(t) to minimize H is that

$$\frac{\partial H}{\partial u}\,(x(t),u(t),J_x^*,t) = 0 \qquad (1\text{-}4)$$

From (1-3) we have

$$\frac{\partial H}{\partial u}\,(x(t),u(t),J_x^*,t) = R(t)u(t)+B^T(t)J_x^*(x(t),t). \qquad (1\text{-}5)$$

Solving (1-4) and (1-5) for u*(t) gives

$$u^*(t) = -R^{-1}(t)B^T(t)J_x^*(x(t),t). \qquad (1\text{-}6)$$

The minimum cost is of the form:

$$J^*(x(t),t) = \frac{1}{2}\, x^T(t)K(t)x(t)\ , \qquad (1\text{-}7)$$

where K(t) is a real, symmetric, positive-definite matrix that is to be determined. It can be shown that K(t) satisfies the Riccati equation ,

$$\dot{K}(t) = -Q(t)-K(t)A(t)-A^T(t)K(t)+K(t)B(t)R^{-1}(t)B^T(t)K(t), \qquad (1\text{-}8a)$$

with boundary condition ,

$$K(t_f) = H\ . \qquad (1\text{-}8b)$$

Substituting (1-7) into (1-6) yields

$$u^*(t) = -R^{-1}(t)B^T(t)K(t)x(t)$$

$$= -L(t)x(t)\ . \qquad (1\text{-}9)$$

The block diagram for the linear regulator problem is shown in Fig. 1-1 .

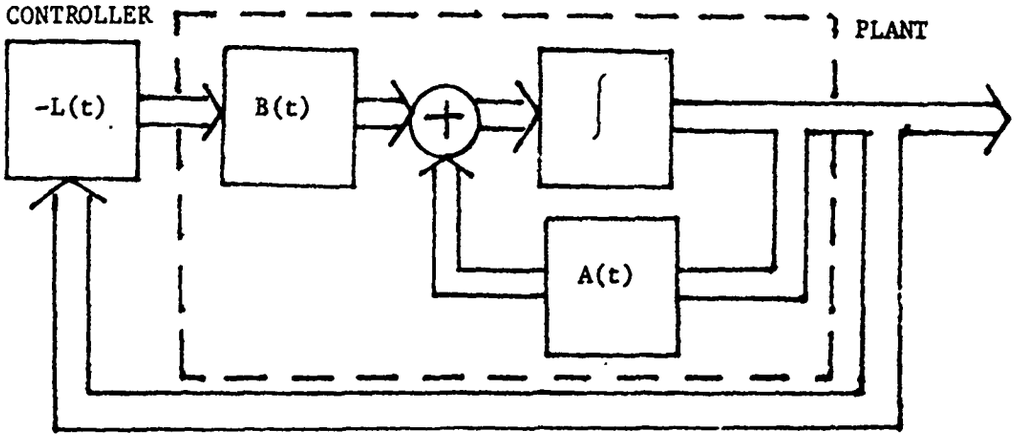


FIGURE 1-1. BLOCK DIAGRAM FOR LINEAR REGULATOR PROBLEM

From (1-7) the linear regulator problem is to maintain the state vector close to a constant without an excessive expenditure of control effort. Such a property can be extended to the linear tracking problem, which keeps the state vector following some specific function. An example of this will be the control system for a radar antenna, whose axis is to be kept aligned from the line of sight to an aircraft flying past with constant angular velocity.

If we apply the optimal control, $u^*(t)$, obtained in (1-9), the optimal trajectory $x^*(t)$ will be :

$$\begin{aligned}\dot{x}^*(t) &= A(x)x^*(t)+B(t)U^*(t) \\ &= [A(t)-B(t)L(t)]x^*(t) \\ &= [A(t)-B(t)R^{-1}(t)B^T(t)K(t)]x^*(t) ,\end{aligned} \tag{1-10}$$

whose poles are eigenvalues of $A(t)+B(t)L(t)$.

For observable control systems, the performance index is often regarded as a weighted measure of the output vector and control vector. Assuming, without loss of generality, that the output vector is

$$y(t) = c(t)x(t) , \tag{1-11}$$

and the quadratic cost function is given by

$$J = \frac{1}{2} y^T(t_f)\hat{H}y(t_f)+ \frac{1}{2} \int_{t_0}^{t_f} [y^T(\lambda)\hat{Q}(\lambda)y(\lambda)+u^T(\lambda)R(\lambda)u(\lambda)]d\lambda ; \tag{1-12}$$

after substitution of (1-11) in (1-12), this yields

$$J = \frac{1}{2} x^T(t_f)C^T(t_f)\hat{H}C(t_f)x(t_f)$$

$$= \frac{1}{2}\int_{t_f}^{t_f} [x^T(\lambda)C^T(\lambda)\hat{Q}(\lambda)C(\lambda)x(\lambda)+u^T(\lambda)R(\lambda)u(\lambda)]d\lambda \ . \qquad (1\text{-}13)$$

Note that by choosing

$$H = C^T(t_f)\hat{H}C(t_f) \quad \text{and} \qquad (1\text{-}14a)$$

$$Q(t) = C^T(t)\hat{Q}(t)C(t) \ , \qquad (1\text{-}14b)$$

(1-13) and (1-2b) are exactly the same. The choice of $\hat{H}$, $\hat{Q}$ and R in (1-12) determines a relative weighting of the various terms. $\hat{Q}$, $\hat{H}$ must be real, symmetric, positive semidefinite matrices and R must be a real symmetric positive definite matrix. Once the designer has specified $\hat{Q}$, $\hat{H}$ and R, representing different weightings in (1-13), the optimal closed-loop system will be

$$\dot{x}^*(t) = [A(t)-B(t)L(t)]x^*(t) \qquad (1\text{-}15a)$$

$$y(t) = Cx(t) \ . \qquad (1\text{-}15b)$$

If the resulting transient response is unsatisfactory, the designer may alter the weighting matrices, $\hat{Q}$ and R, and try again. The use of an optimal observer to realize the optimal control law is discussed in [2].

1.3 Kalman Filter Problem

In Section 1 we considered only systems which were determin-

istic, in the sense that all inputs could be specified exactly, and all outputs could be measured with unlimited precision. These assumptions are mathematically convenient and have led to many powerful and useful theoretical developments. In practice, of course, they cannot always be satisfied. Input and output transducers are subject to unpredictable fluctuations and disturbances, and communication channels are corrupted by all manner of interference. Such uncertainties are present in all physical systems and are usually referred to by the term, noise. In some cases the noise is inconsequential, and a deterministic analysis will suffice. In others, however, the effect of the noise is too great to be ignored and it must be modeled explicitly. The process of analysis and design of these systems needs to be modified. The most commonly used model for this purpose is the stochastic system model.

Stochastic control theory was developed during the Second World War to synthesize fire control systems and radar tracking systems. The propounder of filtering and prediction theory (Wiener-Kolmogorov theory [3]) plays a very important role in the solution of stochastic optimal control problem. Its disadvantage is that it requires the solution of an integral equation (the Wiener-Hopf equation). In realistic problems the Wiener-Hopf equation seldom has analytical solutions, and it is not easy to solve the equation numerically. Nevertheless, the use of the digital computer for both analysis and synthesis has profoundly influenced the development of the theory. Kalman and Bucy [4], [5] made it possible to solve prediction and filtering problems recursively, which is ideally suitable for digital computers. The results of Kalman and Bucy can be applied, not only to the stationary processes, but also to

nonstationary processes. Using the Kalman-Bucy theory, the covariance of the estimation error is governed by a Riccati equation. The Kalman gain (or optimal gain) can be obtained by solving an initial value problem for the Riccati equation which is similar to the one encountered in the optimal control of a deterministic system with quadratic performance index as discussed in Section 1. The state estimation problem and the linear quadratic control problem are, in fact, mathematical duals. This result is of great interest from both the theoretical and the practical points of view. If one of the problems is solved, we can easily obtain the solution of the other by invoking this duality (see Chapter III).

Now consider a stochastic linear system of the following form :

$$\dot{x}(t) = Fx(t)+D\omega(t) \tag{1-16a}$$

$$d(t) = Hx(t)+v(t) , \tag{1-16b}$$

where $F\epsilon R^{q\times q}$, $D\epsilon R^{q\times \ell}$, $H\epsilon R^{p\times q}$, $x\epsilon R^{q}$, $\omega\epsilon R^{\ell}$, $v\epsilon R^{p}$ and $d\epsilon R^{p}$. $\omega(t)$ is called the input noise, $v(t)$ is called the output noise, they are assumed to have zero means and to be white, and

$$E[\omega(t)] = 0 \tag{1-17a}$$

$$E[v(t)] = 0 \tag{1-17b}$$

$$E[\omega(t)\omega^{T}(\tau)] = Q\delta(t-\tau) \tag{1-17c}$$

$$E[v(t)v^{T}(\tau)] = R\delta(t-\tau) \tag{1-17d}$$

$$E[\omega(t)v^T(\tau)] = 0 \qquad \text{(for all } t,\tau\text{)} . \qquad (1\text{-}17e)$$

Note that the differential equation in (1-16) is defined only if we accept the notion of continuous-time white noise. In discrete-time systems, white noise is well defined,and the problem does not arise.

The initial state of the system (1-16) at time,t_0, is usually assumed to be a random vector, $x(t_0) = x_0$, with mean, $E(x_0)$, and covariance :

$$P_0 = E\{[x_0-E(x_0)][x_0-E(x_0)]^T\} , \qquad (1\text{-}18)$$

which is also assumed to be uncorrelated with the noise processes ω and v.

Consider now the problem of estimating the state x(t) of (1-16) at time $t>t_0$, using the noisy measurement data $\{d(t'): t_0 \leq t' < t\}$. To determine an estimate,$\hat{x}(t)$,of x(t),it is common to form the state error vector:

$$\tilde{x}(t) = x(t)-\hat{x}(t), \qquad (1\text{-}19)$$

and then minimize the mean-square error ,

$$E[a^T\tilde{x}(t)]^2 = E[a^T\tilde{x}(t)\tilde{x}^T(t)a]$$

$$= a^T E[\tilde{x}(t)\tilde{x}^T(t)]a , \qquad (1\text{-}20)$$

where $a^T\tilde{x}(t)$ represents any linear combination of the state variables.

It is assumed that

$$\hat{x}(t_0) = E[x(t_0)] \quad . \tag{1-21}$$

The covariance of the estimation error is defined as

$$p(t) = E\{[\tilde{x}(t)-E(\tilde{x}(t)][\tilde{x}(t)-E(\tilde{x}(t)]^T\} \ , \tag{1-22}$$

and the estimator model is given by

$$\dot{\hat{x}}(t) = F\hat{x}(t)+K(t)[d(t)-H\hat{x}(t)] \ . \tag{1-23}$$

Subtracting (1-23) from (1-16a) yields the differential equation for the state estimator ,

$$\begin{aligned}\dot{\tilde{x}}(t) &= F\tilde{x}(t)+D\omega(t)-K(t)[d(t)-H\hat{x}(t)] \\ &= F\tilde{x}(t)-K(t)[Hx(t)+v(t)-H\hat{x}(t)]+D\omega(t) \\ &= [F-K(t)H]\tilde{x}(t)+D\omega(t)-K(t)v(t) \ .\end{aligned} \tag{1-24}$$

From (1-19) and (1-21) it can be shown that

$$E[\tilde{x}(t)] = 0 \ . \tag{1-25}$$

Therefore, (1-22) and (1-20) may be rewritten as

$$p(t) = E[\tilde{x}(t)\tilde{x}^T(t)] \tag{1-26}$$

$$E[a^T\tilde{x}(t)]^2 = a^T p(t) a \,. \tag{1-27}$$

It has been proved [6] that by choosing the gain parameter as

$$K(t) = -p(t)H^T R^{-1} \,, \tag{1-28}$$

the optimal estimation error covariance, p(t), is the symmetric, semi-definite solution of a nonlinear, time-varying matrix differential equation known as a Riccati equation,

$$\dot{p}(t) = Fp(t)+p(t)F^T+DQD^T-p(t)H^T R^{-1}Hp(t) \tag{1-29}$$

$$p(t_0) = P_0 \,. \tag{1-30}$$

P_0 is the covariance matrix of the initial state x_0 and is given in (1-18).

The estimate $\hat{x}(t)$ is unbiased since its averaged error (1-25) is zero, and it is optimal in the sense that at each time, t, its mean-square error is smaller than that achieved by any other linear estimator. If we also make the fairly common assumption that the initial state and the two noise processes satisfy Gaussian (or normal) probability distributions, then the mean-square error is less than that achieved by any other estimator, linear or nonlinear.

The block diagram for a stochastic state estimator problem is shown in Fig. 1-2. Further discussions may be found in reference [7],

[8], and [9].

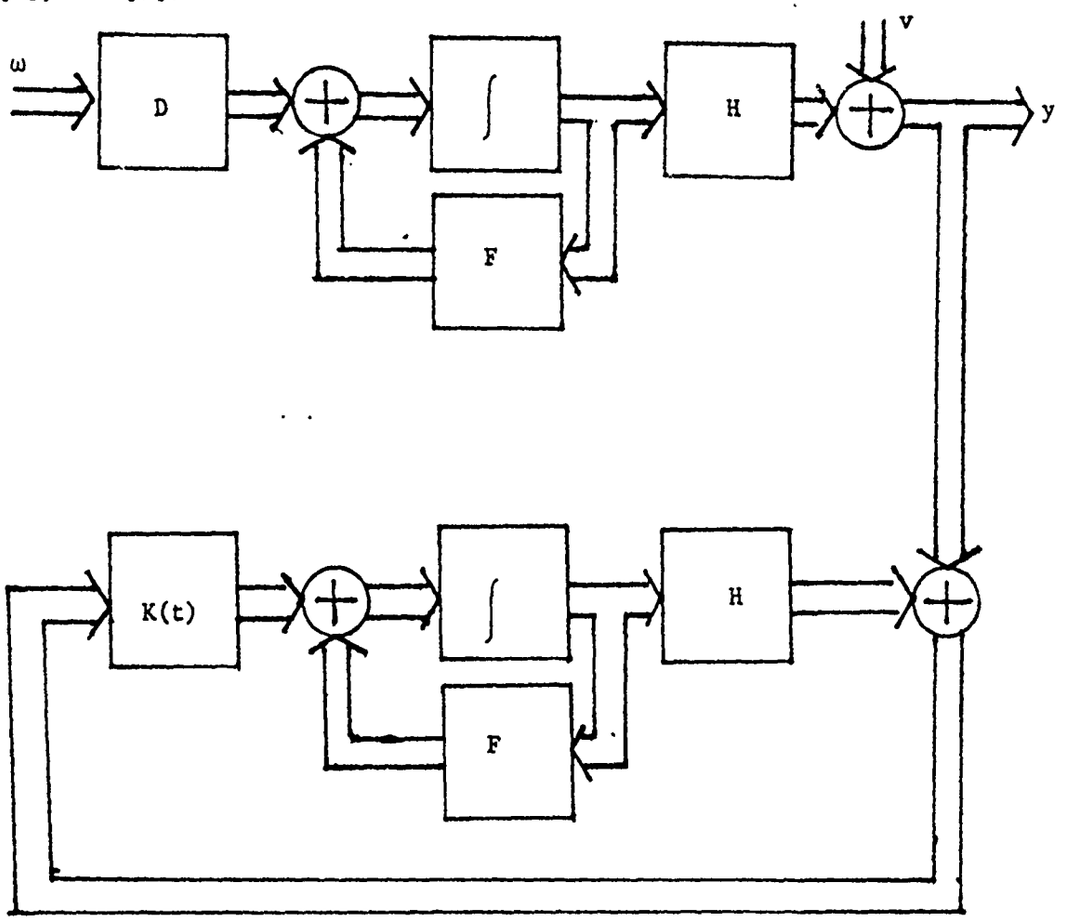


FIGURE 1-2. BLOCK DIAGRAM FOR STOCHASTIC STATE ESTIMATION PROBLEM

CHAPTER II

DISCRETIZATION OF CONTINUOUS-TIME SYSTEM MODEL

2.1 Reasons for Discretization

The accurate description of most practical systems often requires high-order,continuous-time state equations. As a result, the simulation, realization and design of these systems will need to find an explicit solution of differential equations. For a linear, time-invariant system, it is possible to find an analytic solution. However, if the solution is required at many points (e.g., for graph plotting) and if the state vector is at all large, it is exceedingly laborious if done manually. A simpler and much more efficient way to compute the solution is to convert the continuous-time system equation into discrete-time system equations which can be easily implemented by using a digital computer or a microprocessor. Yet finding an exact discrete-time state equation representation is impractical for a large system. Approximation is often used to reduce the computational burden.

2.2 Continuous-Time System Model and Discrete-Time System Model

Consider the system represented by the continuous-time state equation:

$$\dot{x}(t) = Ax(t)+Bu(t) \tag{2-1a}$$

$$x(0) = x_0 \ . \tag{2-1b}$$

The exact solution of (2-1) will be

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-\lambda)}Bu(\lambda)\alpha\lambda \,, \tag{2-2}$$

where $e^{At} = \Phi(t)$ is called the state transition matrix of the system.

For practical consideration [10], we are interested in staircase inputs, or

$$u(t) = u(kT) \triangleq u(k) \tag{2-3}$$

for $k = 0,1,2,3,\ldots,$

and T = a sampling period

with $kT \leq t \leq (k+1)T$.

Substituting (2-3) into (2-2), we find

$$x(k) = \Phi^k(T)x(0) + \sum_{j=0}^{k-1} \Phi(k-j-1)Lu(j) \,, \tag{2-4}$$

where

$x(kT) \triangleq x(k)$

$x(kT+T) \triangleq x(k+1)$

$\Phi(kT-jT+T) \triangleq \Phi(k-j-1)$

$\Phi^k(T) \triangleq [\Phi(T)]^k$ = the continuous-time state transition matrix

and

$$\Phi(T) = e^{AT} = \sum_{j=0}^{\infty} \frac{1}{j!} (AT)^j \quad . \tag{2-5}$$

By letting $\alpha = t-\lambda$, the L-matrix is

$$L = \int_0^T e^{A\alpha} B d\alpha = T \sum_{j=0}^{\infty} \frac{1}{(j+1)!} (AT)^j B$$

$$= [e^{AT}-I]A^{-1}B \, . \tag{2-6}$$

For ease in implementation and manipulation, we are interested in representing a continuous-time state equation by a discrete-time state equation:

$$x^*(k+1) = Dx^*(k)+Eu(k) \tag{2-7a}$$

$$x^*(0) = x(0) \, , \tag{2-7b}$$

where

$$x^*(kT) \triangleq x^*(k) \simeq x(kT)$$

$$x^*(kT+T) \triangleq x^*(k+1) \simeq \dot{x}(kT) \, .$$

The solution of (2-7) is :

$$x^*(k) = D^k x(0) + \sum_{j=0}^{k-1} D^{k-j-1} Eu(j) \, . \tag{2-8}$$

Comparing (2-4) and (2-8), we maintain that $x^*(k)$ will be equal to $x(k)$

if we choose

$$D = \Phi(T) = e^{AT} \tag{2-9a}$$

and $$E = [\Phi(T)-I]A^{-1}B. \tag{2-9b}$$

D is defined as the discrete-time state transition matrix.

2.3 Transition Matrix Approximation

From (2-5) we know that $\Phi(T)$ is an infinite series whose exact value is difficult to obtain when the dimension of A matrix is high. Approximated representation is thus required. A natural question is: how accurately can we approximate $\Phi(T)$? One popular method is to truncate the infinite series, i.e.,

$$\Phi a(T) = \sum_{j=0}^{k} \frac{(AT)^j}{j!} \tag{2-10}$$

When k = 1,2,3,4,5, (2-10) becomes

$$\Phi a(T) = I+AT \tag{2-11a}$$

$$= I+AT + \frac{1}{2!}(AT)^2 \tag{2-11b}$$

$$= I+AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 \tag{2-11c}$$

$$= I+AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \frac{1}{4!}(AT)^4 \tag{2-11d}$$

$$= I+AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \frac{1}{4!}(AT)^4 + \frac{1}{5!}(AT)^5 \tag{2-11e}$$

If k is sufficiently large, a satisfactory approximation may be obtained. However, the approximation error may become serious if the higher order terms in the infinite series have a greater influence on the evaluation of $\Phi(T)$. Such a situation may occur when the number of terms or the sampling period is not properly chosen. This kind of shortcoming can be complemented by the geometric series technique [11]. Now rewrite $\Phi(T)$ as

$$\begin{aligned}\Phi(T) &= e^{AT} \\ &= I+AT + \frac{1}{2!}(AT)^2 +\ldots+ \frac{1}{j!}(AT)^j + \frac{1}{(j+1)!}(AT)^{j+1} \\ &+ \frac{1}{(j+2)!}(AT)^{j+2} + \frac{1}{(j+3)!}(AT)^{j+3} +\ldots+ \frac{1}{(j+n)!}(AT)^{j+n} \\ &+ \ldots \qquad . \end{aligned} \tag{2-12}$$

Keeping the first (j+1) terms in the series of (2-12) and approximating the other terms of the series by a geometric series with a weighting factor $(1/(j^n \cdot j!))$ for the term $(AT)^{j+n}$ rather than $1/(j+n)!$ $(=1/(j+n)(j+n-1)\ldots(j+1)\cdot j!)$ for the same term, we obtain a more accurate model:

$$\Phi_b(T) = \sum_{i=0}^{j-1} \frac{(AT)^i}{i!} + \sum_{i=j}^{\infty} \frac{(AT)^i}{j^{(i-j)}\cdot(j!)} \tag{2-13a}$$

$$= \sum_{i=0}^{j-1} \frac{(AT)^i}{i!} + \frac{(AT)^j}{j!}\left[I - \frac{1}{j}(AT)\right]^{-1} \tag{2-13b}$$

$$= \left[I - \frac{1}{j}(AT)\right]^{-1}\left[I + \sum_{i=1}^{j-1} \frac{(j-i)}{(j)\cdot(i!)}(AT)^i\right] \tag{2-13c}$$

$$= D_{bj}\,, \tag{2-13d}$$

for $T < j/||A||$

$j = 1,2,3,\ldots$.

Note that the second summation term in (2-13a) is a geometric series. The subscript of D_b in (2-13d) indicates that the value of the factor j is to be used in the infinite series. For each D_{bj}, the corresponding E_{bj} can be attained from (2-9b). The approximated modes of D_{bj} and E_{bj} for j = 1,2,3,4,5 are listed as follows:

$$D_{b1} = (I-AT)^{-1} \tag{2-14a}$$

$$D_{b2} = (I-\tfrac{1}{2}AT)^{-1}(I+\tfrac{1}{2}AT) \tag{2-14b}$$

$$D_{b3} = (I-\tfrac{1}{3}AT)^{-1}(I+\tfrac{2}{3}AT+\tfrac{1}{6}(AT)^2) \tag{2-14c}$$

$$D_{b4} = (I-\tfrac{1}{4}AT)^{-1}(I+\tfrac{3}{4}AT+\tfrac{1}{4}(AT)^2+\tfrac{1}{24}(AT)^3) \tag{2-14d}$$

$$D_{b5} = (I-\tfrac{1}{5}AT)^{-1}(I+\tfrac{4}{5}AT+\tfrac{3}{10}(AT)^2+\tfrac{1}{15}(AT)^3+\tfrac{1}{120}(AT)^4), \tag{2-14e}$$

and

$$E_{b1} = T(I-AT)^{-1}B \tag{2-15a}$$

$$E_{b2} = T(I-\tfrac{1}{2}AT)^{-1}B \tag{2-15b}$$

$$E_{b3} = T(I-\tfrac{1}{3}AT)^{-1}(I+\tfrac{1}{6}AT)B \tag{2-15c}$$

$$E_{b4} = T(I-\tfrac{1}{4}AT)^{-1}(I+\tfrac{1}{4}AT+\tfrac{1}{24}(AT)^2)B \tag{2-15d}$$

$$E_{b5} = T(I-\tfrac{1}{5}AT)^{-1}(I+\tfrac{3}{10}AT+\tfrac{1}{15}(AT)^2+\tfrac{1}{120}(AT)^3)B. \tag{2-15e}$$

Given the continuous-time system matrix A and input matrix B, the approximated discrete-time system matrix D_{bj} and input matrix E_{bj} can be calculated by using (2-14) and (2-15). The roundoff errors between the exact mode and the approximated mode increase as $||A||T$ increases. In order to control these errors, we use a scaling and squaring technique. An alternative form of (2-9a) is:

$$\Phi(T) = e^{AT} = (e^{AT_i})^i, \tag{2-16a}$$

where $T_i \triangleq T/i \qquad i = 1,2,3,\ldots$. (2-16b)

From (2-5), we get

$$\Phi(T) = [\sum_{j=0}^{\infty} \frac{1}{j!} (AT_i)^j]^i \tag{2-17a}$$

$$= [\sum_{j=0}^{\infty} \frac{1}{j!} (\frac{AT}{i})^j]^i \tag{2-17b}$$

$$= [\sum_{j=0}^{\infty} \frac{1}{(i^j)\cdot(j!)} (AT)^j]^i \qquad \text{for } i = 1,2,3,\ldots . \tag{2-17c}$$

The infinite series inside the bracket of (2-17c) can be approximated by either truncating the series as in (2-10) or applying a geometric series approach as in (2-13). Consider first the case of truncating the infinite series:

$$\Phi_c(T) = [\sum_{j=0}^{k} \frac{1}{(i^j)\cdot(j!)} (AT)^j]^i \tag{2-18a}$$

$$= [I+\frac{1}{i}(AT)]^i \qquad \text{when } k = 1 \tag{2-18b}$$

$$= [I+\frac{1}{i}(AT)+\frac{1}{2!\cdot(i)^2}(AT)^2]^i \qquad \text{when } k = 2 \tag{2-18c}$$

$$= \ldots\ldots .$$

Of course, the more terms of the Taylor series that are taken, the better the approximation will be. The effect of scaling can be easily seen by considering the case of $i = 1$ and $i = 2$ in (2-18b) and (2-18c). When $i = 1$ (i.e., no scaling is used),

$$\Phi_c(T) = I+AT \tag{2-19a}$$

$$= I+AT+\frac{1}{2!}(AT)^2 \tag{2-19b}$$

$$= I+AT+\frac{1}{2!}(AT)^2+\frac{1}{3!}(AT)^3 ; \tag{2-19c}$$

when $i = 2$ (i.e., scaling is used),

$$\Phi_c(T) = I+AT+\frac{1}{4}(AT)^2 \tag{2-20a}$$

$$= I+AT+\frac{1}{2}(AT)^2+\frac{1}{8}(AT)^3+\frac{1}{64}(AT)^4 \tag{2-20b}$$

$$= I+AT+\frac{1}{2}(AT)^2+\frac{1}{6}(AT)^3+\frac{7}{192}(AT)^4+\frac{1}{192}(AT)^5+\frac{1}{2304}(AT)^6. \tag{2-20c}$$

Comparing (2-19a) and (2-20a), we observe that both retain the first two dominant terms of the Taylor series, but (2-20a) has an extra term, $(AT)^2/4$, which is an approximation of the third term, $(AT)^2/2!$, of the

Taylor series. Therefore, (2-20a) is a better approximation than (2-19a). In the same fashion, comparing (2-19b) and (2-20b), we find that both equations have the first three terms of the Taylor series in (2-5), but (2-20b) provides two more important terms, $(AT)^3/8$ and $(AT)^4/64$, which are the approximations of the respective fourth and fifth terms, ($(AT)^3/6$ and $(AT)^4/24$), of (2-5). From the above comparisons we conclude that (2-20) gives better approximations than (2-19) does.

If we use a geometric series approach to find the value of the series in the bracket of (2-17), $\Phi(T)$ becomes

$$\Phi_d(T) = [\sum_{\ell=0}^{j-1} \frac{(AT)^\ell}{(i^\ell)\cdot(\ell!)} + \sum_{\ell=j}^{\infty} \frac{(AT)^\ell}{(i^\ell)\cdot(j^{\ell-j})\cdot(j!)}]^i \qquad (2\text{-}21a)$$

$$= [\sum_{\ell=0}^{j-1} \frac{(AT)^\ell}{(i^\ell)\cdot(\ell!)} + \frac{(AT)^j}{(i^j)\cdot(j!)} (I - \frac{1}{j\cdot i} AT)^{-1}]^i \qquad (2\text{-}21b)$$

$$= \{[I - \frac{1}{j\cdot i} AT]^{-1}[I + \sum_{\ell=1}^{j-1} \frac{(j-\ell)}{(i^\ell)\cdot(j)\cdot(\ell!)} (AT)^\ell]\}^i, \qquad (2\text{-}21c)$$

$$\text{for } T < i\cdot j/||A|| \qquad (2\text{-}21d)$$

$$i = 1,2,3,\ldots$$

$$j = 1,2,3,\ldots \quad ,$$

where $||A||$ is a matrix norm and $[I - \frac{1}{j\cdot i} (AT)]^{-1}$ is a generalized geometric series. Note that when $i = 1$, $[I - \frac{1}{j} (AT)]$ is a geometric series as that in (2-13b). The approximated discrete transition matrix D and

input matrix E, for i = 1 and j = 1,2,3,4,5 are listed in Eqs. (2-14) and (2-15). When i = 2, the D's and E's matrices for j = 1,2,3,4,5 will be:

$$D_{d1} = [I-AT+ \frac{1}{4} (AT)^2]^{-1} \tag{2-22a}$$

$$D_{d2} = [I- \frac{1}{2} AT+ \frac{1}{16} (AT)^2]^{-1}[I+ \frac{1}{2} AT+ \frac{1}{16} (AT)^2] \tag{2-22b}$$

$$D_{d3} = [I- \frac{1}{3} AT+ \frac{1}{36} (AT)^2]^{-1}[I+ \frac{2}{3} AT+ \frac{7}{36} (AT)^2+ \frac{1}{36} (AT)^3+ \frac{1}{576} (AT)^4] \tag{2-22c}$$

$$D_{d4} = [I- \frac{1}{4} AT+ \frac{1}{64} (AT)^2]^{-1}[I+ \frac{3}{4} AT+ \frac{17}{64} (AT)^2+ \frac{3}{64} (AT)^3+ \frac{1}{128}(AT)^4 + \frac{1}{1536} (AT)^5+ \frac{1}{36864} (AT)^6] \tag{2-22d}$$

$$D_{d5} = [I- \frac{1}{5} AT+ \frac{1}{100} (AT)^2]^{-1}*[I+ \frac{4}{5} AT+ \frac{31}{100} (AT)^2+ \frac{3}{200} (AT)^3 + \frac{59}{4800} (AT)^4+ \frac{1}{800} (AT)^5+ \frac{17}{115200} (AT)^6+ \frac{1}{115200} (AT)^7 + \frac{1}{3686400} (AT)^8] , \tag{2-22e}$$

and

$$E_{d1} = T[I- \frac{1}{4} AT+ \frac{1}{4} (AT)^2]^{-1}(I- \frac{1}{4} AT)B \tag{2-23a}$$

$$E_{d2} = T[I- \frac{1}{2} AT+ \frac{1}{16} (AT)^2]^{-1}B \tag{2-23b}$$

$$E_{d3} = T[I- \frac{1}{3} AT+ \frac{1}{36} (AT)^2]^{-1}[I+ \frac{1}{6} AT+ \frac{1}{36} (AT)^2+ \frac{1}{576} (AT)^3]B \tag{2-23c}$$

$$E_{d4} = T[I- \frac{1}{4} AT+ \frac{1}{64} (AT)^2]^{-1}[I+ \frac{1}{4} AT+ \frac{3}{64} (AT)^2+ \frac{1}{128} (AT)^3 + \frac{1}{1536} (AT)^4+ \frac{1}{36864} (AT)^5]B \tag{2-23d}$$

$$E_{d5} = T[I - \frac{1}{5}AT + \frac{1}{100}(AT)^2]^{-1}[I + \frac{3}{10}AT + \frac{3}{200}(AT)^2 + \frac{59}{4800}(AT)^3 + \frac{1}{800}(AT)^4 + \frac{17}{115200}(AT)^5 + \frac{1}{115200}(AT)^6 + \frac{1}{3686400}(AT)]B$$

The boundary conditions for the choice of convergence of the sampling period T in (2-13d) and (2-21d) are different from each other by a factor of i. Therefore, by using the generalized geometric series, one can use a larger sampling period as long as we use a larger scaling factor i. This is an important property which makes possible the on-line calculation by applying microcomputers or microprocessors, because small computers exchange the price with the speed and capacity.
The approximation $\Phi_d(T)$ in (2-21), not only retains the first (j+1) dominant terms of the Taylor series in (2-5), but also approximates the rest of that infinite series. Therefore the accuracy of $\Phi_d(T)$ is much better than that of $\Phi_c(T)$ in (2-18), which preserves the first few terms (depending on how many terms we choose in the bracket), approximates some terms thereafter, and truncates all higher order terms. For example, when i = 1 and j = 2,

$$\Phi_c(T) = I + AT + \frac{1}{2}(AT)^2 \qquad (2\text{-}24a)$$

$$\Phi_d(T) = (I - \frac{1}{2}AT)^{-1}(I + \frac{1}{2}AT)$$

$$= I + AT + \frac{1}{2}(AT)^2 + \frac{1}{2^2 \cdot 1}(AT)^3 + \frac{1}{2^3 \cdot 1}(AT)^4 + \frac{1}{2^4 \cdot 1}(AT)^5 + \ldots \qquad (2\text{-}24b)$$

When i = 2 and j = 2,

$$\Phi_c(T) = [I+\frac{1}{2}AT+\frac{1}{8}(AT)^2]^2$$

$$= I+AT+\frac{1}{2}(AT)^2+\frac{1}{2^2\cdot 2}(AT)^3+\frac{1}{2^3\cdot 2^3}(AT)^4 \qquad (2\text{-}25a)$$

$$\Phi_d(T) = [(I-\frac{1}{4}AT)^{-1}(I+\frac{1}{4}AT)]^2$$

$$= I+AT+\frac{1}{2}(AT)^2+\frac{1}{2^2\cdot(\frac{4}{3})}(AT)^3+\frac{1}{2^3\cdot 2}(AT)^4+\frac{1}{2^4\cdot(\frac{16}{5})}(AT)^5+\ldots . \qquad (2\text{-}25b)$$

Rewriting (2-5) for the comparison of (2-24) and (2-25) with the exact discrete transition matrix, $\Phi(T)$, we have:

$$\Phi(T) = I+AT+\frac{1}{2}(AT)^2+\frac{1}{2^2\cdot(\frac{3}{2})}(AT)^3+\frac{1}{2^3\cdot 3}(AT)^4+\frac{1}{2^4\cdot(\frac{15}{2})}(AT)^5+\ldots . \qquad (2\text{-}26)$$

It is obvious from (2-24)~(2-26) that the first three dominant terms of all five equations are identical, and the coefficients of the remaining terms in (2-24) and (2-25) compared with (2-26) give the conclusion that a better discrete-time state transition matrix can be constructed by using the generalized geometric series rather than using the scaled truncating method.

The matrix $\Phi(T)$ can also be obtained by modifying e^{AT} as follows:

$$\Phi(T) = e^{AT} = (e^{-\frac{1}{2}AT})^{-1}(e^{\frac{1}{2}AT}), \qquad (2\text{-}27)$$

where $e^{-\frac{1}{2}AT}$ and $e^{\frac{1}{2}AT}$ can be acquired by using the truncating model in (2-10), or the geometric series model in (2-13). The corresponding discrete system matrices D and E are:

$$D \simeq \Phi_e(T) = \left[\sum_{i=0}^{j} \frac{(-1)^i (AT)^i}{(2^i)\cdot(i!)}\right]^{-1}\left[\sum_{i=0}^{j} \frac{(AT)^i}{(2^i)\cdot(i!)}\right] \qquad (2\text{-}28a)$$

$$E \simeq [\Phi_e(T)-I]A^{-1}B = \left[\sum_{i=0}^{j} \frac{(-1)^i\cdot(AT)^i}{(2^i)\cdot(i!)}\right]^{-1}* \left\{\sum_{i=0}^{INT[(j-1)/2]} \frac{(AT)^{2i}}{(2^{2i})\cdot[(2i+1)!]}\right\}B, \qquad (2\text{-}28b)$$

where INT[(j-1/2] represents the integer part of the real number (j-1)/2 and

$$D \simeq \Phi_f(T) = \left\{\sum_{i=0}^{j} \frac{(-1)^i[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT)^i\right\}^{-1}* \left\{\sum_{i=0}^{j} \frac{[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT)^i\right\} \qquad (2\text{-}29a)$$

$$E \simeq [\Phi_f(T)-I]A^{-1}B = T\left\{\sum_{i=0}^{j} \frac{(-1)^i\cdot[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT)^i\right\}^{-1}* \left\{\sum_{i=1}^{j} \frac{[1-(-1)^i][j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT)^{i-1}\right\}B. \qquad (2\text{-}29b)$$

For j = 1,2,3,4,5, the approximate models are:

$$D_{e1} = (I-\tfrac{1}{2}AT)^{-1}(I+\tfrac{1}{2}AT) \qquad (2\text{-}30a)$$

$$D_{e2} = [I-\tfrac{1}{2}AT+\tfrac{1}{8}(AT)^2]^{-1}[I+\tfrac{1}{2}AT+\tfrac{1}{8}(AT)^2] \qquad (2\text{-}30b)$$

$$D_{e3} = [I-\tfrac{1}{2}AT+\tfrac{1}{8}(AT)^2-\tfrac{1}{48}(AT)^3]^{-1}[I+\tfrac{1}{2}AT+\tfrac{1}{8}(AT)^2+\tfrac{1}{48}(AT)^3] \qquad (2\text{-}30c)$$

$$D_{e4} = [I - \frac{1}{2}AT + \frac{1}{8}(AT)^2 - \frac{1}{48}(AT)^3 + \frac{1}{384}(AT)^4]^{-1}[I + \frac{1}{2}AT + \frac{1}{8}(AT)^2 + \frac{1}{48}(AT)^3 + \frac{1}{384}(AT)^4] \quad (2\text{-}30d)$$

$$D_{e5} = [I - \frac{1}{2}AT + \frac{1}{8}(AT)^2 - \frac{1}{48}(AT)^3 + \frac{1}{384}(AT)^4 - \frac{1}{3840}(AT)^5]^{-1} * [I + \frac{1}{2}AT + \frac{1}{8}(AT)^2 + \frac{1}{48}(AT)^3 + \frac{1}{384}(AT)^4 + \frac{1}{3840}(AT)^5] \quad (2\text{-}30e)$$

$$E_{e1} = T(I - \frac{1}{2}AT)^{-1}B \quad (2\text{-}31a)$$

$$E_{e2} = T(I - \frac{1}{2}AT + \frac{1}{8}(AT)^2]^{-1}B \quad (2\text{-}31b)$$

$$E_{e3} = T[I - \frac{1}{2}AT + \frac{1}{8}(AT)^2 - \frac{1}{48}(AT)^3]^{-1}[I + \frac{1}{24}(AT)^2]B \quad (2\text{-}31c)$$

$$E_{e4} = T[I - \frac{1}{2}AT + \frac{1}{8}(AT)^2 - \frac{1}{48}(AT)^3 + \frac{1}{384}(AT)^4]^{-1}[I + \frac{1}{24}(AT)^2]B \quad (2\text{-}31d)$$

$$E_{e5} = T[I - \frac{1}{2}AT + \frac{1}{8}(AT)^2 - \frac{1}{48}(AT)^3 + \frac{1}{384}(AT)^4 - \frac{1}{3840}(AT)^5]^{-1} [I + \frac{1}{24}(AT)^2 + \frac{1}{1920}(AT)^4]B. \quad (2\text{-}31e)$$

Note that (2-30) and (2-31) are obtained from (2-28). As for (2-29), the D's and E's become:

$$D_{f1} = (I - \frac{1}{2}AT)^{-1}(I + \frac{1}{2}AT) \quad (2\text{-}32a)$$

$$D_{f2} = [I - \frac{1}{2}AT + \frac{1}{16}(AT)^2]^{-1}[I + \frac{1}{2}AT + \frac{1}{16}(AT)^2] \quad (2\text{-}32b)$$

$$D_{f3} = [I - \frac{1}{2}AT + \frac{7}{72}(AT)^2 - \frac{1}{144}(AT)^3]^{-1}[I + \frac{1}{2}AT + \frac{7}{72}(AT)^2 + \frac{1}{144}(AT)^3] \quad (2\text{-}32c)$$

$$D_{f4} = [I - \frac{1}{2}AT + \frac{7}{64}(AT)^2 - \frac{5}{384}(AT)^3 + \frac{1}{1536}(AT)^4]^{-1}[I + \frac{1}{2}AT + \frac{7}{64}(AT)^2 + \frac{5}{384}(AT)^3 + \frac{1}{1536}(AT)^4] \quad (2\text{-}32d)$$

$$D_{f5} = [I - \frac{1}{2}AT + \frac{23}{200}(AT)^2 - \frac{19}{1200}(AT)^3 + \frac{13}{9600}(AT)^4 - \frac{1}{19200}(AT)^5]^{-1} * [I + \frac{1}{2}AT + \frac{23}{200}(AT)^2 + \frac{19}{1200}(AT)^3 + \frac{13}{9600}(AT)^4 + \frac{1}{19200}(AT)^5] \quad (2\text{-}32e)$$

and

$$E_{f1} = T(I - \frac{1}{2}AT)^{-1}B \quad (2\text{-}33a)$$

$$E_{f2} = T[I - \frac{1}{2}AT + \frac{1}{16}(AT)^2]^{-1}B \quad (2\text{-}33b)$$

$$E_{f3} = T[I - \frac{1}{2}AT + \frac{7}{72}(AT)^2 - \frac{1}{144}(AT)^3]^{-1}[I + \frac{1}{72}(AT)^2]B \quad (2\text{-}33c)$$

$$E_{f4} = T[I - \frac{1}{2}AT + \frac{7}{64}(AT)^2 - \frac{5}{384}(AT)^3 + \frac{1}{1536}(AT)^4]^{-1}[I + \frac{5}{192}(AT)^2]B \quad (2\text{-}33d)$$

$$E_{f5} = T[I - \frac{1}{2}AT + \frac{23}{200}(AT)^2 - \frac{19}{1200}(AT)^3 + \frac{13}{9600}(AT)^4 - \frac{1}{19200}(AT)^5]^{-1} [I + \frac{19}{600}(AT)^2 + \frac{1}{9600}(AT)^4]B \quad (2\text{-}33e)$$

Comparing (2-30) and (2-32) with (2-11) and (2-14), we conclude that the modified transition matrix in (2-27) gives a better result than the original transition matrix in (2-5). In addition, the modified e^{AT} implies a bilinear representation for the transition matrix D. This bilinearity is useful in solving problems with large and small eigenvalues.

Furthermore, we put forward the best approximation for the state transition matrix in (2-5):

$$\Phi_g(T) = e^{AT} \tag{2-34a}$$

$$= [(e^{\frac{-1}{2} AT_n})^{-1}(e^{\frac{1}{2} AT_n})]^n \tag{2-34b}$$

$$= Q_{jn}^{-n} P_{jn}^{n} , \tag{2-34c}$$

where

$$T_n = T/n \tag{2-34d}$$

$$Q_{jn} = [I- \frac{1}{2\cdot j\cdot n} (AT)][I+ \sum_{i=1}^{j-1} \frac{(-1)^i(j-i)}{(2^i)\cdot(j)\cdot(i!)\cdot(n^i)} (AT)^i] \tag{2-34e}$$

$$P_{jn} = [I+ \frac{1}{2\cdot j\cdot n} (AT)][I+ \sum_{i=1}^{j-1} \frac{(j-i)}{(2^i)\cdot(j)\cdot(i!)\cdot(n^i)} (AT)^i] \tag{2-34f}$$

for $j = 1,2,3,...$

$n = 1,2,3,...$

with $T < (2\cdot j\cdot n)/||A||$. (2-34g)

Equation (2-34) can be regarded as the scaling and squaring model for (2-29a). Hence a larger sampling period can be used, and the accuracy is improved. For convenience, we list some approximants (Φ_{jn} for $j = 1,2,3,4,5$):

$$\Phi_{1n} = [(I- \frac{1}{2n} AT)^{-1}(I+ \frac{1}{2n} AT)]^n \tag{2-35a}$$

$$\Phi_{2n} = \{[I - \frac{1}{2n} AT + \frac{1}{16n^2} (AT)^2]^{-1}[I + \frac{1}{2n} AT + \frac{1}{16n^2} (AT)^2]\}^n \quad (2\text{-}35b)$$

$$\Phi_{3n} = \{[I - \frac{1}{2n} AT + \frac{7}{72n^2} (AT)^2 - \frac{1}{144n^3} (AT)^3]^{-1}[I + \frac{1}{2n} AT + \frac{7}{72n^2} (AT)^2 + \frac{1}{144n^3} (AT)^3]\}^n \quad (2\text{-}35c)$$

$$\Phi_{4n} = \{[I - \frac{1}{2n} AT + \frac{7}{64n^2} (AT)^2 - \frac{5}{384n^3} (AT)^3 + \frac{1}{1536n^4} (AT)^4]^{-1} * [I + \frac{1}{2n} AT + \frac{7}{64n^2} (AT)^2 + \frac{5}{384n^3} (AT)^3 + \frac{1}{1536n^4} (AT)^4]\}^n \quad (2\text{-}35d)$$

$$\Phi_{5n} = \{[I - \frac{1}{2n} AT + \frac{23}{200n^2} (AT)^2 - \frac{19}{1200n^3} (AT)^3 + \frac{13}{9600n^4} (AT)^4 - \frac{1}{19200n^5} (AT)^5]^{-1}[I + \frac{1}{2n} AT + \frac{23}{200n^2} (AT)^2 + \frac{19}{1200n^3} (AT)^3 + \frac{13}{9600n^4} (AT)^4 + \frac{1}{19200n^5} (AT)^5]\}^n . \quad (2\text{-}35e)$$

Note that the coefficients of two matrix polynomials in each Φ_{jn} in (2-35) are identical except for signs. As a result, Φ_{jn} can be evaluated faster than other approximation methods, and the computational error in evaluating the approximate transition matrix may be minimized.

2.4 Summary

The above discussion may be summarized as follows:

1. The exact transformation from continuous-time system (Eq. (2-1)) into discrete-time system (Eq. (2-7)) uses the relation:

$$D = \Phi(T) = e^{AT}$$

$$E = (D-I)A^{-1}B$$

under the assumption that the input is a piecewise constant input (the case of piecewise linear input will be discussed in Chapter IV).

2. Seven different approximations for $\Phi(T)$ are derived. $\Phi_a(T)$, $\Phi_c(T)$ and $\Phi_e(T)$ in (2-10), (2-18) and (2-28a) respectively use truncating method, whereas $\Phi_b(T)$, $\Phi_d(T)$ and $\Phi_f(T)$ in (2-13), (2-21) and (2-29a) respectively use the geometric series approach. Φ_c and Φ_d are obtained by scaling and squaring Φ_a and Φ_b, while Φ_e and Φ_f are found from Φ_a and Φ_b by applying (2-27).

3. $\Phi_g(T)$ is the best approximant of $\Phi(T)$ since it has largest convergent range, minimum computational error, and fastest calculation speed. In addition, the peculiar bilinear matrix expansion format is particularly useful in solving a stiff state-space equation [12] which has both large and small eigenvalues for which the Runge-Kutta fourth-order integration method [13] fails. This is due to the fact that the Runge-Kutta method approximates the Taylor series matrix expansion by taking the first five dominant terms only, whereas $\Phi_g(T)$ uses, not only the first several dominant terms, but also an infinite number of other approximate terms.

4. Other approximation techniques for the transition matrix $\Phi(T)$ can be found in [14].

CHAPTER III

APPROXIMATED LINEAR REGULATOR AND KALMAN FILTER

3.1 Introduction

In Chapter I, a linear regulator problem has been illustrated as one of the optimal control design techniques that have general applications in deterministic systems. A regulator problem is defined as an optimal feedback control system that will drive the states or outputs to the neighborhood of the equilibrium conditions. However, most real dynamic control systems have disturbances and measurement noises. It is not possible, for example, to model a disturbance by an analytical function. The answer to the problem of modeling disturbances is to describe them as stochastic processes. The Kalman filter has been shown to have applications in stochastic control problems [6,15] and is particularly effective in the estimation of system states contaminated by white noise. One of the difficulties in the determination of the Kalman filter is the computational burden encountered in computing the filter error covariance matrix for use in obtaining the gains. Consequently, approximation methods discussed in Chapter II are used to obtain a suboptimal state estimation which can be easily implemented on a low cost minicomputer or microprocessor.

For the deterministic optimal linear regulator problem [Eq. (1-2)], which is the mathematical dual of the optimal stochastic-state estimation problem [Eq. (1-16)], Kleinman [16] et al. have proposed a very elegant approach in solving the suboptimal linear regulator problem by using piecewise-constant gains. Chen and Shiao [17] have devised a Walsh function approach for developing a piecewise-constant gain to

approximate the time-varying Kalman gain, while Rao [18] has improved the computational speed of the Walsh function approach via the block-pulse function technique. In this chapter the generalized geometric series approach and scaling-squaring technique mentioned in Chapter II will be used for developing piecewise-constant gains and piecewise-linear gains for approximations of the optimal gains and Kalman gains in the linear regulator problem and the state estimation problem, respectively. Also, simple and fast algorithms are presented for the implementation of these problems on a computer.

3.2 Time-Varying Optimal Gain

Rewriting the optimal linear regulator system in (1-2) as follows:

$$\dot{x}(t) = Ax(t) + Bu(t); \qquad x(t_0) = x_0 , \tag{3-1}$$

where x(t), u(t), A and B are vectors and matrices of appropriate dimensions. For a finite time t_f, the quadratic loss function,

$$J = \frac{1}{2}\int_{t_0}^{t_f} [x^T(t)Qx(t)+u^T(t)Ru(t)]dt, \tag{3-2a}$$

is minimized using the optimal control law (1-9):

$$u(t) = -L(t)x(t) = -R^{-1}B^T\lambda(t) \tag{3-2b}$$

where the time-varying optimal gain is:

$$L(t) = R^{-1}B^T\hat{p}(t) , \tag{3-2c}$$

and the adjoint state variable $\lambda(t)$ is equal to

$$\lambda(t) = \hat{p}(t)\, x(t) \tag{3-2d}$$

$\hat{p}(t)$ satisfies the following Hamiltonian matrix equation [19]:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\lambda}(t) \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix} \triangleq \hat{M} \begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix}, \tag{3-3a}$$

and the boundary conditions are specified as:

$$\begin{bmatrix} x(t_0) \\ \hat{p}(t_f) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}. \tag{3-3b}$$

Instead of solving the time-varying optimal gain $L(t)$ from a Riccati equation [20] by off-line computations, we can solve a linear two-point-boundary-value problem [19] for the $L(t)$. The procedures are reviewed as follows:

The solution of (3-3) is:

$$\begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix} = \hat{\Phi}(t,t_0) \begin{bmatrix} x(t_0) \\ \lambda(t_0) \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\Phi}_{11}(t,t_0) & \hat{\Phi}_{12}(t,t_0) \\ \hat{\Phi}_{21}(t,t_0) & \hat{\Phi}_{22}(t,t_0) \end{bmatrix} \begin{bmatrix} x(t_0) \\ \lambda(t_0) \end{bmatrix}, \tag{3-4a}$$

where the continuous transition matrix . :

$$\hat{\Phi}(t,t_0) = e^{\hat{M}(t,t_0)}. \tag{3-4b}$$

Partitioning $\hat{\Phi}(t,t_0)$ and substituting $t = t_f$ gives:

$$\lambda(t_f) = \hat{\Phi}_{21}(t_f,t_0)x(t_0)+\hat{\Phi}_{22}(t_f,t_0)\lambda(t_0). \tag{3-5}$$

Using (3-2d) and (3-3b), we find

$$\lambda(t_0) = -\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)x(t_0). \tag{3-6}$$

From (3-4a),

$$\begin{aligned} x(t) &= \hat{\Phi}_{11}(t,t_0)x(t_0)+\hat{\Phi}_{12}(t,t_0)\lambda(t_0) \\ &= [\hat{\Phi}_{11}(t,t_0)-\hat{\Phi}_{12}(t,t_0)\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)]x(t_0) \end{aligned} \tag{3-7a}$$

$$\begin{aligned} \lambda(t) &= \hat{\Phi}_{21}(t,t_0)x(t_0)+\hat{\Phi}_{22}(t,t_0)\lambda(t_0) \\ &= [\hat{\Phi}_{21}(t,t_0)-\hat{\Phi}_{22}(t,t_0)\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)]x(t_0), \end{aligned} \tag{3-7b}$$

therefore

$$\begin{aligned} \lambda(t) = [\hat{\Phi}_{21}(t,t_0)-\hat{\Phi}_{22}(t,t_0)\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)] \cdot \\ [\hat{\Phi}_{11}(t,t_0)-\hat{\Phi}_{12}(t,t_0)\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)]^{-1}\cdot x(t). \end{aligned} \tag{3-8}$$

Hence $\hat{p}(t)$ is found as:

$$\hat{p}(t) = [\hat{\Phi}_{21}(t,t_0)-\hat{\Phi}_{22}(t,t_0)\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)]\cdot$$

$$[\hat{\Phi}_{11}(t,t_0)-\hat{\Phi}_{12}(t,t_0)\hat{\Phi}_{22}^{-1}(t_f,t_0)\hat{\Phi}_{21}(t_f,t_0)]^{-1}. \qquad (3\text{-}9a)$$

The time-varying optimal gain is

$$L(t) = R^{-1}B^T\hat{p}(t), \qquad (3\text{-}9b)$$

and the optimal state trajectory is given in (3-7a).

3.3 Time-Varying Kalman Gain

Furthermore, we review the continuous stochastic-state estimation problem. Consider the linear time-invariant continuous stochastic system given as:

$$\dot{x}(t) = Fx(t) + D\omega(t) \; ; \quad x(t_0) = x_0 \qquad (3\text{-}10a)$$

$$d(t) = Hx(t) + v(t), \qquad (3\text{-}10b)$$

where $x(t)\epsilon R^q$, $d(t)\epsilon R^p$, $F\epsilon R^{q\times q}$, $D\epsilon R^{q\times \ell}$, $H\epsilon R^{p\times q}$, $\omega(t)\epsilon R^{\ell}$, $v(t)\epsilon R^p$; $\omega(t)$ and $v(t)$ are zero mean stationary white noise processes having the properties:

$$E[\omega(t)\omega^T(\tau)] = Q\delta(t-\tau) \qquad Q \geq 0 \qquad (3\text{-}10c)$$

$$E[v(t)v^T(\tau)] = R\delta(t-\tau) \qquad R > 0 \qquad (3\text{-}10d)$$

$$E[\omega(t)v^T(\tau)] = 0 \qquad (3\text{-}10e)$$

The best linear least squares estimate of the state vector is given by the stochastic state equation [Section 2 of Chapter I]:

$$\dot{\hat{x}}(t) = F\hat{x}(t)+K(t)[d(t)-H\hat{x}(t)] \qquad (3\text{-}11a)$$

$$\hat{x}(t_0) = E[x(t_0)], \qquad (3\text{-}11b)$$

where $\hat{x}(t)$ is the estimated state and the Kalman gain is:

$$K(t) = \tilde{P}(t)H^T R^{-1}. \qquad (3\text{-}11c)$$

$\tilde{P}(t)$ is the covariance of the estimation error.

Introduce the vector z defined as the solution of the differential equation:

$$\dot{z}(t) = -F^T z(t)-H^T u(t) \; ; \qquad z(t_0) = z_0, \qquad (3\text{-}12)$$

then the estimation problem is equivalent to the problem of finding a control signal for the dynamical system (3-12) which minimizes the quadratic performance index:

$$J = \frac{1}{2} z^T(t_0)\tilde{p}(t_0)z(t_0) + \frac{1}{2}\int_{t_0}^{t_f} [z^T(t)DQD^T z(t)+u^T(t)Ru(T)]dt. \qquad (3\text{-}13)$$

Consequently, the state estimation problem becomes the mathematical dual of the optimal control problem, and the optimal control law is thus

equal to

$$u(t) = -R^{-1}H\tilde{p}(t)z(t) = -K^{T}(t)z(t). \tag{3-14}$$

Similar to (3-3a), we can form the following Hamiltonian matrix equation.

$$\begin{bmatrix} \dot{z}(t) \\ \dot{\lambda}(t) \end{bmatrix} = \begin{bmatrix} -F^{T} & H^{T}R^{-1}H \\ DQD^{T} & F \end{bmatrix} \begin{bmatrix} z(t) \\ \lambda(t) \end{bmatrix} \triangleq \tilde{M} \begin{bmatrix} z(t) \\ \lambda(t) \end{bmatrix}, \tag{3-15a}$$

with boundary conditions,

$$\lambda(t_0) = \tilde{p}(t_0)z(t_0) \tag{3-15b}$$

$$\tilde{p}(t_0) = E\{[x(t_0)-Ex(t_0)][x(t_0)-Ex(t_0)]^{T}\}, \tag{3-15c}$$

and the adjoint state variable $\lambda(t)$ satisfies

$$\lambda(t) = \tilde{p}(t)\ z(t), \tag{3-16}$$

where $\tilde{p}(t)$ is the covariance matrix.

The solution of (3-15) is:

$$\begin{bmatrix} z(t) \\ \lambda(t) \end{bmatrix} = \tilde{\Phi}(t,t_0) \begin{bmatrix} z(t_0) \\ \lambda(t_0) \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\Phi}_{11}(t,t_0) & \tilde{\Psi}_{12}(t,t_0) \\ \tilde{\Phi}_{21}(t,t_0) & \tilde{\Phi}_{22}(t,t_0) \end{bmatrix} \begin{bmatrix} z(t_0) \\ \lambda(t_0) \end{bmatrix},$$

where the continuous transition matrix is:

$$\tilde{\Phi}(t,t_0) = e^{\tilde{M}(t,t_0)}. \tag{3-17b}$$

After partitioning the matrix in (3-17a) and using the relationship in (3-15) and (3-16), we obtain:

$$\tilde{P}(t) = [\tilde{\Phi}_{21}(t,t_0)+\tilde{\Phi}_{22}(t,t_0)\tilde{P}(t_0)][\Phi_{11}(t,t_0)+\Phi_{12}(t,t_0)\tilde{P}(t_0)]^{-1}. \tag{3-18a}$$

Note that (3-18a) is different from (3-9a). When t_f in (3-13) is a finite time, the time-varying Kalman gain is

$$K(t) = \tilde{P}(t)H^T R^{-1} \tag{3-18b}$$

Substituting (3-18) into (3-11) yields the optimally estimated state, $\hat{x}(t)$.

3.4 Optimal Regulator and Kalman Filter Approximation

In Chapter II, several approximation methods for the transition matrix are discussed. The newest and probably the best one among them is shown in (2-34). Since both (3-4b) and (3-17b) can be treated as transition matrices, and due to the duality of stochastic-state estimation and deterministic optimal regulator, we can extend (2-34) to construct a piecewise-constant gain and a piecewise time-varying gain for approximating L(t) in (3-9b) and K(t) in (3-18b). As it shall be seen later, the proposed method improves the accuracy and computational speed of the existing methods [17,18], and the approximate gains obtained can be

implemented on low-cost microprocessors or minicomputers for on-line suboptimal control and approximate estimation [21,22] of a wide class of systems.

If the exact transition matrices, $\hat{\Phi}(t,t_0)$ in (3-4) and $\tilde{\Phi}(t,t_0)$ in (3-17), can be obtained by off-line computation, the exact time-varying optimal gain L(t) in (3-9b) and the exact time-varying Kalman gain K(t) in (3-18b) can be determined for optimal control and estimation. Moreover, off-line computation can be achieved by using a huge and expensive digital computer, but it may not be practical to implement it on a small and low cost minicomputer or microprocessor because of its slow speed and limited capacity. For this reason, approximants are often determined and implemented on a mini/micro computer for on-line suboptimal control and approximate estimation. Chen and Hsiao [17] approximated $\hat{\Phi}(t,t_0)$, but not $\tilde{\Phi}(t,t_0)$, via a Walsh function approach, while Rao [18] approximated $\hat{\Phi}(t,t_0)$ via a block-pulse function approach. Considering practical engineering constraints, we choose the modified geometric series approach with scaling and squaring, which is a class of Pade approximation method [14], to approximate both $\hat{\Phi}(t,t_0)$ and $\tilde{\Phi}(t,t_0)$. This method will improve the accuracy and computational speed of the existing methods [17,18].

A general continuous-time state equation,

$$\dot{Y}(t) = MY(t) \; ; \qquad Y(t_0) = Y(0), \tag{3-19a}$$

is used to represent (3-3a) and (3-15a). The solution of (3-19a) is :

$$Y(t) = e^{Mt}\, Y(0) = \Phi(t)\, Y(0) \; , \tag{3-19b}$$

where $\Phi(t)[= e^{Mt}]$ is a continuous transition matrix. To use the recursive feature of a discrete-time formulation and programmable microprocessors or minicomputers, the continuous state equation in (3-19a) is often converted to an equivalent discrete-time model as:

$$Y(KT+T) = GY(KT) \; ; \qquad Y(t_0) = Y(0). \tag{3-20a}$$

Thus, the discrete-time solution can be rapidly determined as:

$$Y(KT) = G^K Y(0), \tag{3-20b}$$

$$\text{where} \quad G = e^{MT} \triangleq \Phi_d(T) \tag{3-20c}$$

$$\Phi(t) = \Phi_d^K(t) = [e^{MT}]^K = G^K \tag{3-20d}$$

$$t = KT, \quad K = 0,1,2,... \tag{3-20e}$$

$T(= t/K)$ is the sampling period and $\Phi_d(T)$ is a discrete transition matrix. If off-line computations of $\Phi_d(T)$ are not available or not desired (for example, the self-tuning control problem [23] and the adaptive control problem [21,22]), and the on-line suboptimal control or approximated estimation using a microprocessor or a minicomputer is permissible, then the $\Phi_d(T)$ is often approximated by a matrix polynomial or a rational matrix polynomial. Once the approximation of G has been determined, the approximate discretized solution of (3-19) becomes:

$$Y_d(KT) = G^K Y_d(0) \; ; \qquad Y_d(0) = Y(0), \tag{3-21a}$$

where

$$G \simeq \Phi_d(T) \tag{3-21b}$$

$$Y_d(KT) \simeq Y(t) \qquad \text{at} \quad t = KT \tag{3-21c}$$

Using (2-34), we get the best approximation of G:

$$G = e^{MT} = \{[e^{\frac{-1}{2} MT_n}]^{-1}[e^{\frac{1}{2} MT_n}]\}^n$$

$$\simeq Q_{jn}^{-n} P_{jn}^{n} \triangleq G_{jn} \tag{3-22a}$$

for $j = 1,2,\ldots$

$n = 1,2,\ldots$,

where $T_n \triangleq T/n$

$$Q_{jn} \triangleq [I_{2q} - \frac{1}{(2)(j)(n)} MT][I_{2q} + \sum_{i=1}^{j-1} \frac{(-1)^i (j-i)(MT)^i}{(2^i)(j)(i!)(n^i)}] \tag{3-22b}$$

$$P_{jn} \triangleq [I_{2q} + \frac{1}{(2)(j)(n)} MT][I_{2q} + \sum_{i=1}^{j-1} \frac{(j-i)(MT)^i}{(2^i)(j)(i!)(n^i)}] \tag{3-22c}$$

$$T < (2jn)/||M|| . \tag{3-22d}$$

I_{2q} is a $2q \times 2q$ identity matrix, $||M||$ is a matrix norm of M and the rational matrix polynomial $[I_{2q} - \frac{1}{2jn} MT]^{-n}$ is a geometric series. Now we shall investigate the accuracy and computational speed of the proposed method compared with other existing methods. When $n = 1$, G_{jn}

(for j = 1,2,...) in (3-22a) are the approximations of the e^{MT} obtained by taking the first (j+2) dominant terms and an infinite number of the other approximated terms of the Taylor series matrix expansion. For example, when n = 1 and j = 1, G_{11} is given by:

$$G_{11} = [I_{2q} - \frac{1}{2} MT]^{-1}[I_{2q} + \frac{1}{2} MT] \tag{3-23a}$$

$$= I_{2q} + MT + \frac{1}{2!} (MT)^2 + \sum_{j=3}^{\infty} \frac{1}{2^{j-1}} (MT)^j \tag{3-23b}$$

$$T < 2/||M||, \tag{3-23c}$$

the exact Taylor series matrix expansion is

$$G = e^{MT} = I_{2q} + MT + \frac{1}{2!} (MT)^2 + \sum_{j=3}^{\infty} \frac{1}{j!} (MT)^j \tag{3-24}$$

Observe that the first three dominant terms in both (3-23b) and (3-24) are identical and the remaining terms differ by their weighting factor $1/(2^{j-1})$ in (3-23b) and $1/(j!)$ in (3-24). Shieh [24] et al. have shown that the discrete-time solution in (3-21), having G_{11}(=G) in (3-23), is identical to the approximated solution of (3-19) obtained by using Walsh function approach [25] and the block-pulse function approach [12]. Since G_{11} is a special case of G_{jn} in (3-22), the implication is that the gains designed via the existing methods [17,18] are the special cases of the modified geometric series with the scaling and squaring method.

For practical implementation of the designed approximate optimal gains and Kalman gains on a microprocessor, which needs a large

sampling period due to its slower operation, a more sophisticated equivalent model (G_{jn} an j>1 and/or n>1) is required. For instance, if letting j=1 and n>1 in (3-22), we have:

$$G_{1n} = Q_{1n}^{-n} P_{1n}^{n}$$

$$= \{[I_{2q} - \frac{1}{2n} MT]^{-1}[I_{2q} + \frac{1}{2n} MT]\}^{n}$$

$$= \{[I_{2q} - \frac{1}{2} MT_n]^{-1}[I_{2q} + \frac{1}{2} MT_n]\}^{n} \qquad (3\text{-}25a)$$

$$T < 2n/||M|| \qquad (3\text{-}25b)$$

or $$T_n < 2/||M||. \qquad (3\text{-}25c)$$

Comparing (3-23c) and (3-25c), we observe that the sampling period T in (3-23c) has been reduced to T_n (= T/n, n>1) in (3-25c). As a result, the accuracy of the approximation G_{1n} in (3-25a) is better than that of G_{11} in (3-23a). Thus, the proposed method has significantly reformed the accuracy of the existing methods [17,18] for evaluating $\hat{\Phi}(t,t_0)$ and L(t). Also, the range of convergence of the geometric series in (3-25b) has been increased to n times that in (3-23b). From this observation we can conclude that a larger sampling period can be used if a more sophisticated model is chosen. Furthermore, we see that both G_{11} and G_{1n} are in the bilinear matrix expansion format which can be easily applied in solving a stiff state-space equation [12]. Equations, (3-22b) and (3-22c), can be rewritten as follows:

$$Q_{jn} \triangleq [I_{2q} + \sum_{i=1}^{j} \frac{(-1)^i (j^2 - i^2 + i)(MT)^i}{(2^i)\cdot(j^2)\cdot(i!)\cdot(n^i)}] \qquad (3\text{-}26a)$$

$$P_{jn} \triangleq [I_{2q} + \sum_{i=1}^{j} \frac{(j^2-i^2+i)(MT)^i}{(2^i)\cdot(j^2)\cdot(i!)(n^i)}]. \qquad (3\text{-}26b)$$

Note that each term in Q_{jn} and P_{jn} is equal except for signs. As a result, G_{jn} can be evaluated faster than other classes of the Pade approximation, and the computational errors in evaluating the approximate transition matrix may be minimized. These improvements are another reason for choosing a geometric-series approach (a class of Pade approximation approach [14]) for transition matrices approximation.

The G_{jn} matrices for j = 1,2,3,4,5 are shown in (2-35). Substituting any one of them into (3-19) yields an approximate discretized transition matrix $\Phi(t)$ at t = KT. Using this $\Phi(KT)$, we can determine the approximate discretized L(t) in (3-9b) and K(t) in (3-18b). The desired piecewise-constant, approximate optimal gain ($\hat{L}_{pc}(t)$) and the piecewise-constant, approximate Kalman gain ($\hat{K}_{pc}(t)$), derived from a rectangular rule for continuous system control and estimation, are:

$$\hat{L}_{pc}(t) = L(jT) \simeq L(t) \; ; \quad jT \leq t < (j+1)T, \quad j = 0,1,2,\ldots,m-1 \qquad (3\text{-}27a)$$

and

$$\hat{K}_{pc}(t) = K(jT) \simeq K(t) \; ; \quad jT \leq t < (j+1)T, \quad j = 0,1,2,\ldots,m-1, \qquad (3\text{-}27b)$$

where $m(=t_f/T)$ is the number of sub-intervals with sampling period T and a finite time t_f of interest. If a trapezoidal rule is applied, the piecewise-constant gains are:

$$L_{pc}(t) = \frac{1}{2}\,[L(jT+T)+L(jT)] \simeq L(t), \quad jT \le t < (j+1)T,\ j = 0,1,2,\ldots,m-1 \tag{3-28a}$$

and

$$K_{pc}(t) = \frac{1}{2}\,[K(jT+T)+K(jT)] \simeq K(t) \quad jT \le t < (j+1)T,\ j = 0,1,2,\ldots,m-1. \tag{3-28b}$$

To improve the accuracy of the approximate gains in (3-27) and (3-28), we use new piecewise time-varying gains such as:

$$L_{pt}(t) = L(jT) + \frac{1}{T}\,[L(jT+T)-L(jT)](t-jT) \simeq L(t) \tag{3-29a}$$

and

$$K_{pt}(t) = K(jT) + \frac{1}{T}\,[K(jT+T)-K(jT)](t-jT) \simeq K(t), \tag{3-29b}$$

where $jT \le t < (j+1)T$ and $j = 0,1,2,\ldots,m-1$.

To reduce the number of piecewise gains [$L_{pc}(t)$, $L_{pt}(t)$, $K_{pc}(t)$ and $K_{pt}(t)$] from m to ℓ, we further approximate the piecewise gains. The average gain of $\hat{L}_{pc}(t)$ in (3-27a) between the sampling period $T^*(=nT$, $n>1)$ is:

$$\hat{L}^{+}_{pc}(t) = \frac{1}{n} \sum_{i=jn}^{(j+1)n-1} L(iT); \quad jT^* \le t < (j+1)T^*,\ j = 0,1,2,\ldots,\ell-1, \tag{3-30a}$$

where $t_f = mT$,

$\ell = m/n$ is the number of intervals with sampling period

$T^*(=nT)$ and n is the number of subintervals in each interval.

The average gain of $L_{pc}(t)$ in (3-28a) becomes:

$$L^+_{pc}(t) = \frac{1}{2n} \sum_{i=jn}^{(j+1)n-1} [L(iT+T)+L(iT)]; \quad jT^* \leq t < (j+1)T^*, \quad j = 0,1,\ldots,\ell-1. \tag{3-30b}$$

Moreover, the average gain of $L_{pt}(t)$ in (3-29a) is:

$$L^+_{pt}(t) = L(jT^*) + \frac{1}{T^*} [f(jT^*+T^*)-L(jT^*)](t-jT^*), \tag{3-30c}$$

where $$f(jT^*+T^*) = 2L^+_{pc}(t)-L(jT^*) \tag{3-30d}$$

and $$jT^* \leq t < (j+1)T^*, \qquad j = 0,1,2,\ldots,\ell-1.$$

In the same fashion, the average approximate Kalman gains between the sampling period $T^*(=nT, n>1)$ become:

$$\hat{K}^+_{pc}(t) = \frac{1}{n} \sum_{i=jn}^{(j+1)n-1} K(iT) \tag{3-31a}$$

$$K^+_{pc}(t) = \frac{1}{2n} \sum_{i=jn}^{(j+1)n-1} [K(iT+T)+K(iT)] \tag{3-31b}$$

and $$K^+_{pt}(t) = K(jT^*) + \frac{1}{T^*} [g(jT^*+T^*)-K(jT^*)] \quad (t-jT^*), \tag{3-31c}$$

where $$g(jT^*+T^*) = 2K^+_{pc}(t)-K(jT^*) \tag{3-31d}$$

$$jT^* \leq t < (j+1)T^*, \qquad j = 0,1,2,\ldots,\ell-1.$$

The developed approximate optimal gains and approximate Kalman gains in (3-27) to (3-31) can be implemented on the programmable digital controller for on-line suboptimal control and approximate estimation of a system.

3.5 Examples

Now we shall investigate one deterministic problem and one stochastic problem to see how the proposed method improves the result.

Example 1. Deterministic Control Problem

Since the state estimation problem is the dual of the deterministic control problem, we can use Kleinman's deterministic system [16], which has been solved by using piecewise-constant gains, as an illustrative example to test the aforementioned method.

The dynamic equation is:

$$\dot{x}(t) = Ax(t) + Bu(t)$$

$$= \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 2 \\ 2 \\ -1 \end{bmatrix} u(t). \qquad (3\text{-}32a)$$

The initial conditions are:

$$x(0) = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}. \qquad (3\text{-}32b)$$

The quadratic lost function is :

$$J = \frac{1}{2}\int_{t_0}^{t_f} (x^T Q x + u^T R u)\,dt$$

$$= \frac{1}{2}\int_0^2 \left\{ x^T \begin{bmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} x + u^T \cdot 2 \cdot u \right\} dt$$

$$= \int_0^2 [(x_1 - x_2)^2 + u^2]\,dt. \qquad (3\text{-}33)$$

The corresponding state and costate equations are:

$$\dot{Y} = \begin{bmatrix} \dot{x} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix}$$

$$= M \begin{bmatrix} x \\ \lambda \end{bmatrix} = MY, \qquad (3\text{-}34)$$

where $Y(t) = [x(t)\ \lambda(t)]^T$, $\lambda(t) = P(t)\,x(t)$ and $P(t_f) = 0$. After substitution, the M matrix becomes :

$$M = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 0 & 0 & 4 & 4 & -2 \\ 0 & 0 & 2 & 4 & 4 & -2 \\ 0 & -2 & 0 & -2 & -2 & 1 \\ -2 & 2 & 0 & 1 & 0 & 0 \\ 2 & -2 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & -2 & 0 \end{bmatrix}. \qquad (3\text{-}35)$$

The linear control law $u^*(t)$ is thus equal to [Eq. (3-9b)]:

$$u^*(t) = -L(t)x(t)$$

$$= -R^{-1}B^T P(t)x(t). \qquad (3\text{-}36)$$

The desired optimal state trajectory satisfies:

$$\dot{x}^*(t) = (A-BL)x^*(t) \; ; \quad x^*(0) = x(0). \qquad (3\text{-}37)$$

Using (3-19)-(3-22), we obtain the approximate discrete state model of the continuous state model in (3-34) as:

$$Y^*[(j+1)T] = GY^*(jT) \; ; \quad Y^*(0) = Y(0) \qquad (3\text{-}38)$$

where $j = 0,1,2,\ldots,m-1$ and $m = t_f/T$.

The discrete system matrix G can be expressed by various approximations obtained in Chapter II. Here we use the following four sets of G for comparison:

$$G_1 = [I - \tfrac{1}{2}MT]^{-1}[I + \tfrac{1}{2}MT] \text{ as n=1 in (2-35a)} \qquad (3\text{-}39a)$$

$$G_2 = \{[I - \tfrac{1}{4}MT]^{-1}[I + \tfrac{1}{4}MT]\}^2 \text{ as n=2 in (2-35a)} \qquad (3\text{-}39b)$$

$$G_3 = \{[I - \tfrac{1}{4}MT + \tfrac{1}{64}(MT)^2]^{-1}[I + \tfrac{1}{4}MT + \tfrac{1}{64}(MT)^2]\}^2$$

$$\text{as n=2 in (2-35b)} \qquad (3\text{-}39c)$$

$$G_4 = [I - \frac{1}{2}MT + \frac{7}{72}(MT)^2 - \frac{1}{144}(MT)^3]^{-1}[I + \frac{1}{2}MT + \frac{7}{72}(MT)^2 + \frac{1}{144}(MT)^3]$$

as n=1 in (2-35c). (3-39d)

Multiplying G_i (i = 1,2,3,4) in (3-39) j times (j = 1,2,...,m) gives:

$$G_i^j = \begin{bmatrix} \Phi_{i11}(jT) & \Phi_{i12}(jT) \\ \Phi_{i21}(jT) & \Phi_{i22}(jT) \end{bmatrix} = \Phi_i(jT). \quad (3\text{-}40)$$

The corresponding discrete feedback gains $L_i(jT)$ become:

$$L_i(jT) = R^{-1}B^T P_i(jT), \quad (3\text{-}41a)$$

where

$$P_i(jT) = [\Phi_{i21}(jT) - \Phi_{i22}(jT)\Phi_{i22}^{-1}(MT)\Phi_{i21}(MT)]*$$
$$[\Phi_{i11}(jT) - \Phi_{i12}(jT)\Phi_{i22}^{-1}(MT)\Phi_{i21}(MT)]^{-1}. \quad (3\text{-}41b)$$

Since the $L_i(jT)$ in this example is a 3×1 dimensional vector, we denote each element of this column vector as $L_{1i}(jT)$, $L_{2i}(jT)$ and $L_{3i}(jT)$. Tables 3-1, 3-2 and 3-3 show the optimal gains obtained for $L_{1i}(t)$, $L_{2i}(t)$ and $L_{3i}(t)$ respectively, where t = jT, j = 0,1,2,...,m and m=64 or 8 or 4. From Tables 3-1 to 3-3 we observe that when a larger number of intervals, m, is used, a better approximate result is achieved. In addition, a better approximate model (G_i as i>>1) results in a better

i	t	0	1/4	1/2	3/4	1	5/4	3/2	7/4	2
	m=64	0.43700	0.43509	0.43479	0.44082	0.41037	0.28485	0.12089	0.02022	-10^{-5}
$L_{11}(t)$	m=8	0.43857	0.43387	0.43474	0.44426	0.41467	0.29060	0.13058	0.02904	10^{-5}
	m=4	0.43662		0.43701		0.42632		0.15842		10^{-5}
	m=64	0.43698	0.43510	0.43479	0.44078	0.41032	0.28478	0.12077	0.02012	-10^{-6}
$L_{12}(t)$	m=8	0.43747	0.43487	0.43476	0.44162	0.41143	0.28624	0.12323	0.02237	10^{-5}
	m=4	0.43857		0.43474		0.41467		0.13058		10^{-5}
	m=64	0.43697	0.43510	0.43479	0.44077	0.41031	0.28477	0.12074	0.02010	10^{-5}
$L_{13}(t)$	m=8	0.43710	0.43505	0.43478	0.44097	0.41059	0.28513	0.12136	0.02066	-10^{-5}
	m=4	0.43747		0.43476		0.41142		0.12323		-10^{-5}
	m=64	0.43697	0.43510	0.43479	0.44076	0.41030	0.28476	0.12073	0.02008	-10^{-5}
$L_{14}(t)$	m=8	0.43697	0.43511	0.43479	0.44077	0.41031	0.28477	0.12074	0.02010	10^{-6}
	m=4	0.43705		0.43482		0.41039		0.12080		10^{-6}

TABLE 3-1. THE DISCRETE-TIME DATA OF $L_{1i}(jT)$

i	t	0	1/4	1/2	3/4	1	5/4	3/2	7/4	2
	m=64	0.15321	0.15643	0.15593	0.06795	-0.11695	-0.21311	-0.12605	-0.02254	2×10^{-6}
$L_{21}(t)$	m=8	0.13635	0.14687	0.14300	0.03968	-0.15009	-0.23174	-0.13645	-0.03075	-2×10^{-6}
	m=4	0.09763		0.09628		-0.23235		-0.15841		-1×10^{-6}
	m=64	0.15342	0.15655	0.15608	0.06828	-0.11653	-0.21287	-0.12592	-0.02244	10^{-6}
$L_{22}(t)$	m=8	0.14909	0.15402	0.15294	0.06132	-0.12509	-0.21789	-0.12872	-0.02461	-2×10^{-6}
	m=4	0.13635		0.14300		-0.15009		-0.13645		-2×10^{-6}
	m=64	0.15347	0.15658	0.15611	0.06836	-0.11643	-0.21281	-0.12588	-0.02242	-2×10^{-6}
$L_{23}(t)$	m=8	0.15238	0.15594	0.15534	0.06663	-0.11859	-0.21408	-0.12659	-0.02296	2×10^{-6}
	m=4	0.14909		0.15294		-0.12509		-0.12872		-10^{6}
	m=64	0.15349	0.15659	0.15613	0.06839	-0.11640	-0.21278	-0.12587	-0.02243	10^{-6}
$L_{24}(t)$	m=8	0.15346	0.15657	0.15611	0.06836	-0.11644	-0.21282	-0.12590	-0.02243	-10^{-6}
	m=4	0.15308		0.15584		-0.11706		-0.12624		10^{-6}

TABLE 3-2. THE DISCRETE-TIME DATA OF $L_{2i}(jT)$

i	t	0	1/4	1/2	3/4	1	5/4	3/2	7/4	2
	m=64	-0.86400	-0.79956	-0.79326	-0.82766	-0.73191	-0.41140	-0.11476	-0.00875	-10^{-6}
$L_{31}(t)$	m=8	-0.82917	-0.76352	-0.76903	-0.80681	-0.69820	-0.38021	-0.10336	-0.00769	2×10^{-6}
	m=4	-0.71304		-0.71491		-0.60110		-0.07921		10^{-6}
	m=64	-0.86439	-0.79999	-0.79357	-0.82792	-0.73232	-0.41179	-0.11491	-0.00876	-4×10^{-6}
$L_{32}(t)$	m=8	-0.85606	-0.79087	-0.78720	-0.82263	-0.72386	-0.40371	-0.11186	-0.00844	2×10^{-6}
	m=4	-0.82917		-0.76903		-0.69820		-0.10336		3×10^{-6}
	m=64	-0.86449	-0.80010	-0.79365	-0.82798	-0.73242	-0.41188	-0.11495	-0.00877	2×10^{-6}
$L_{33}(t)$	m=8	-0.86243	-0.79782	-0.79204	-0.82665	-0.73030	-0.40985	-0.11417	-0.00869	-3×10^{-6}
	m=4	-0.85606		-0.78720		-0.72386		-0.11186		2×10^{-6}
	m=64	-0.86452	-0.80014	-0.79368	-0.82800	-0.73245	-0.41192	-0.11496	-0.00877	-2×10^{-6}
$L_{34}(t)$	m=8	-0.86448	-0.80009	-0.79363	-0.82797	-0.73240	-0.41189	-0.11494	-0.00876	10^{-6}
	m=4	-0.86396		-0.79300		-0.73198		-0.11470		2×10^{-6}

TABLE 3-3. THE DISCRETE-TIME DATA OF $L_{3i}(jT)$

discrete optimal gain. This implies that a larger sampling period can be used if a more sophisticated model is chosen.

Using the different L_i's obtained from (3-41a) for m=64, 32, 16, 8, 4 or 2, we find the L_{pc}'s in (3-28). From (3-37) the discrete optimal state trajectory is attained. Consequently, the performance indices (defined as J_{ci}) for the various $L_i(t)$ can be calculated by using the trapezoidal rule and are listed in Table 3-4.

From Table 3-4 we observe that a more sophisticated model gives a more precise performance index.

To reduce the number of piecewise gains, an averaging technique as shown in (3-30) and (3-31) is applied. Various performance indices, J_{ci}^{+} and J_{ki}^{+} with m=64 and K = 2,4,8,16 or 32 in (3-30b) and (3-30c), are listed in Tables 3-5 and 3-6, respectively.

Comparing the data in Tables 3-5 and 3-6 we observe that the performance indices in Table 3-6 which use ℓ piecewise time-varying gains are better than those in Table 3-5, which use ℓ piecewise-constant gains. Also, comparing the performance indices in Tables 3-4 through 3-6, we conclude that the performance indices in Table 3-5 and Table 3-6 (which use ℓ piecewise gains where ℓ = 32,16,...,2) are slightly larger than those in the first column of Table 3-4 (which use m=64 piecewise-constant gains). However, a smaller number of simplified piecewise gains is used in Table 3-5 and 3-6. Note that $\ell<m$. Furthermore, the performance indices obtained from ℓ simplified piecewise gains in Tables 3-5 and 3-6 are better than those obtained from the same number of piecewise gains in Table 3-4.

i^m	64	32	16	8	4	2
J_{c1}	1.68907	1.69223	1.70617	1.76519	2.01198	2.87484
J_{c2}	1.68850	1.68995	1.69708	1.72875	1.85897	2.31682
J_{c3}	1.68836	1.68938	1.69482	1.71977	1.82204	2.16286
J_{c4}	1.68831	1.68919	1.69407	1.71686	1.81097	2.13606

TABLE 3-4. THE PERFORMANCE INDICES OBTAINED BY USING $L_{pc}(t)$

i^ℓ	32	16	8	4	2
J^+_{c1}	1.68909	1.68919	1.68956	1.69150	1.69945
J^+_{c2}	1.68853	1.68862	1.68899	1.69093	1.69886
J^+_{c3}	1.68838	1.68848	1.68885	1.69079	1.69871
J^+_{c4}	1.68834	1.68843	1.68880	1.69074	1.69866

TABLE 3-5. THE PERFORMANCE INDICES OBTAINED BY USING $L^+_{pc}(t)$ WITH m=64

i^ℓ	32	16	8	4	2
J^+_{t1}	1.68907	1.68907	1.68913	1.68993	1.69915
J^+_{t2}	1.68850	1.68850	1.68856	1.68935	1.69855
J^+_{t3}	1.68836	1.68836	1.68842	1.68921	1.69840
J^+_{t4}	1.68831	1.68831	1.68837	1.68916	1.69835

TABLE 3-6. THE PERFORMANCE INDICES OBTAINED BY USING $L^+_{pt}(t)$ WITH m=64

Example 2. Stochastic Control Problem

Consider a one-dimensional tracking problem

$$\dot{x}(t) = Fx(t) + D\omega(t)$$

$$= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \omega(t);\ x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{3-42a}$$

$$d(t) = Hx(t) + v(t)$$

$$= [1\ 0]\ x(t) + v(t), \tag{3-42b}$$

where $x_1(t)$ is a noise-free position function and $x_2(t)$ is a constant velocity corrupted by a Gaussian white noise with covariance $Q = 0.1$. The radar detects the position, $x_1(t)$, and is corrupted by a Gaussian white noise with covariance $R = 0.5$. The velocity state disturbance, $\omega(t)$, and the measurement noise, $v(t)$, satisfies:

$$E[\omega(t),\omega^T(\tau)] = Q\delta(t-\tau) = 0.1\ \delta(t-\tau) \tag{3-43a}$$

$$E[v(t)v^T(\tau)] = R\delta(t-\tau) = 0.5\ \delta(t-\tau) \tag{3-43b}$$

$$E[\omega(t)v^T(\tau)] = 0. \tag{3-43c}$$

The initial condition is :

$$x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3-43d}$$

From Eqs. (3-15) through (3-17), we have

$$M = \begin{bmatrix} -F^T & H^T R^{-1} H \\ DQD^T & F \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 2 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.1 & 0 & 0 \end{bmatrix} \qquad (3\text{-}44a)$$

and

$$\Phi(T) = e^{MT}, \qquad (3\text{-}44b)$$

The estimated initial conditions are chosen as :

$$\hat{x}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \qquad (3\text{-}45a)$$

The corresponding covariance matrix becomes :

$$p(0) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \qquad (3\text{-}45b)$$

By applying (3-19) and (3-20) and using the four approximation modes in (3-39), we can derive the Kalman gains $K_{1i}(t)$ and $K_{2i}(t)$ for $i = 1,2,3,4$, where $t = jT$, $j = 0,1,2,\ldots,m$ and $m = 200$ or 8 or 4. The result is shown in Tables 3-7 and 3-8 for $K_{1i}(t)$ and $K_{2i}(t)$, respectively. Sur-

i	t	0	5/8	5/4	15/8	5/2	25/8	15/4	35/8	5
	m=200	0	0.68482	1.39549	1.35712	1.17748	1.04460	0.97027	0.93847	0.93140
K_{11}	m=8	0	0.70609	1.44268	1.38289	1.18701	1.04605	0.96823	0.93560	0.92910
	m=4	0		1.60596		1.21929		0.96163		0.92110
	m=200	0	0.68480	1.39544	1.35709	1.17747	1.04460	0.97028	0.93848	0.93140
K_{12}	m=8	0	0.69003	1.40693	1.36339	1.17980	1.04495	0.96977	0.93778	0.93080
	m=4	0		1.44268		1.18701		0.96823		0.92910
	m=200	0	0.68480	1.39544	1.35709	1.17747	1.04460	0.97028	0.93848	0.93140
K_{13}	m=8	0	0.69001	1.40692	1.36340	1.17981	1.04496	0.96978	0.93778	0.93084
	m=4	0		1.44245		1.18725		0.96835		0.92910
	m=200	0	0.68479	1.39542	1.35708	1.17746	1.04459	0.97028	0.93848	0.93140
K_{14}	m=8	0	0.68480	1.39542	1.35707	1.17746	1.04459	0.97027	0.93848	0.93140
	m=4	0		1.39547		1.17739		0.97024		0.93140

TABLE 3-7. THE DISCRETE-TIME DATA OF $K_{1i}(jT)$

i	t	0	5/8	5/4	15/8	5/2	25/8	15/4	35/8	5
K_{21}	m=200	0	1.10775	1.15075	0.78287	0.56213	0.46428	0.43075	0.42702	0.43389
	m=8	0	1.14729	1.18902	0.79397	0.56235	0.46130	0.42753	0.42471	0.43273
	m=4	0		1.32405		0.56414		0.41692		0.42877
K_{22}	m=200	0	1.10770	1.15070	0.78285	0.56213	0.46429	0.43076	0.42702	0.43389
	m=8	0	1.11738	1.15999	0.78555	0.56217	0.46355	0.42997	0.42646	0.43361
	m=4	0		1.18902		0.56235		0.42753		0.43273
K_{23}	m=200	0	1.10770	1.15070	0.78285	0.56213	0.46429	0.43076	0.42702	0.43389
	m=8	0	1.11737	1.16001	0.78337	0.56218	0.46356	0.42997	0.42646	0.43361
	m=4	0		1.18924		0.56257		0.42757		0.43271
K_{24}	m=200	0	1.10769	1.15069	0.78285	0.56213	0.46429	0.43076	0.42702	0.43389
	m=8	0	1.10769	1.15068	0.78284	0.56213	0.46429	0.43076	0.42702	0.43389
	m=4	0		1.15061		0.56206		0.43075		0.43390

TABLE 3-8. THE DISCRETE-TIME DATA OF $K_{2i}(jT)$

prisingly, all estimated states obtained by using various modes and various piecewise states give good estimation. In Figures 3-1 and 3-2, the simulation results with G and G_1 in Eq. (3-39a) are plotted. In order to demonstrate the effect of averaging, the states simulated from averaged piecewise-constant Kalman gains and averaged piecewise time-varying gains are also included. Here the number of intervals (=ℓ) and the number of subintervals (=K) are chosen to be 8 and 25.

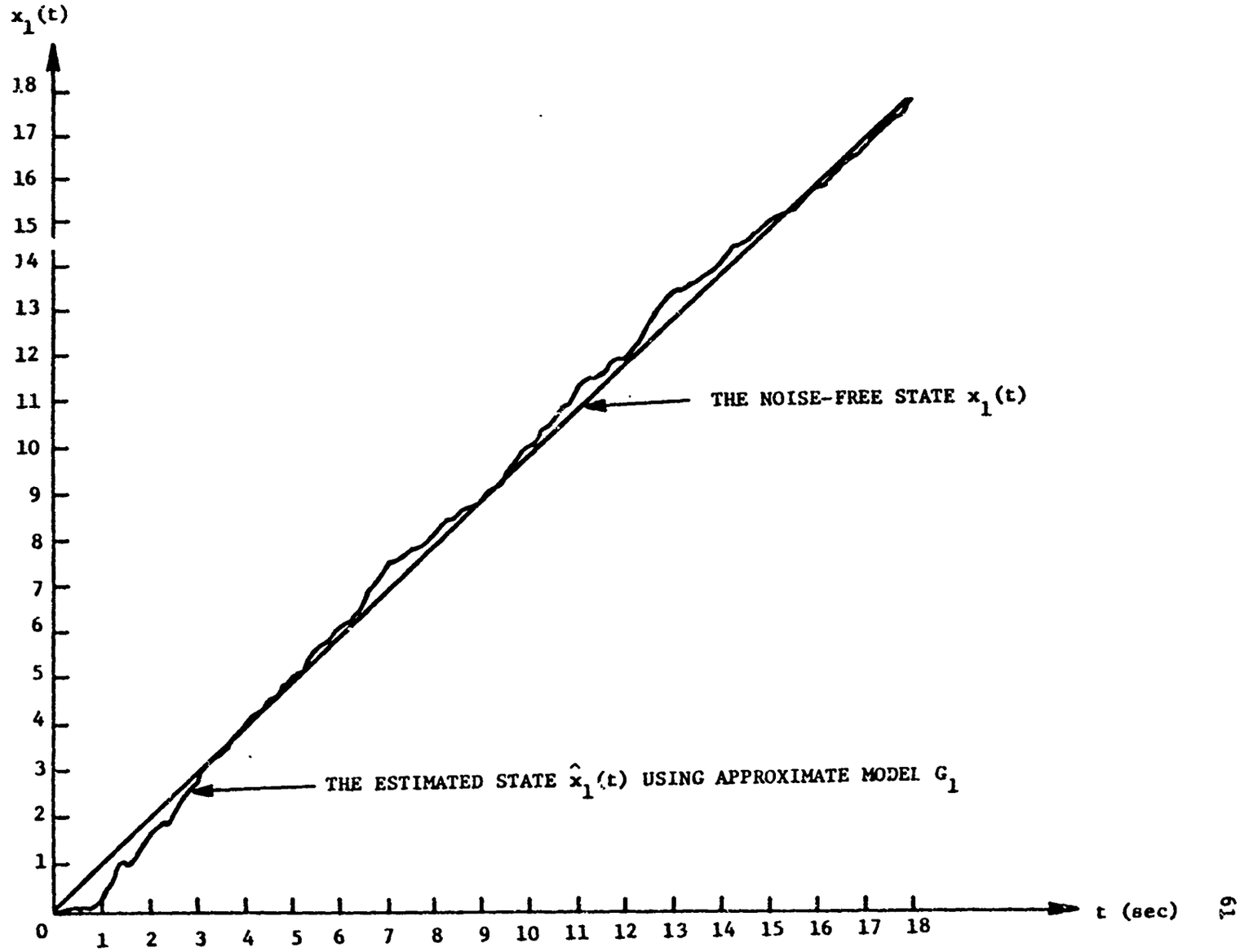


FIGURE 1. THE NOISE-FREE STATE $x_1(t)$ AND THE ESTIMATED STATE $\hat{x}_1(t)$

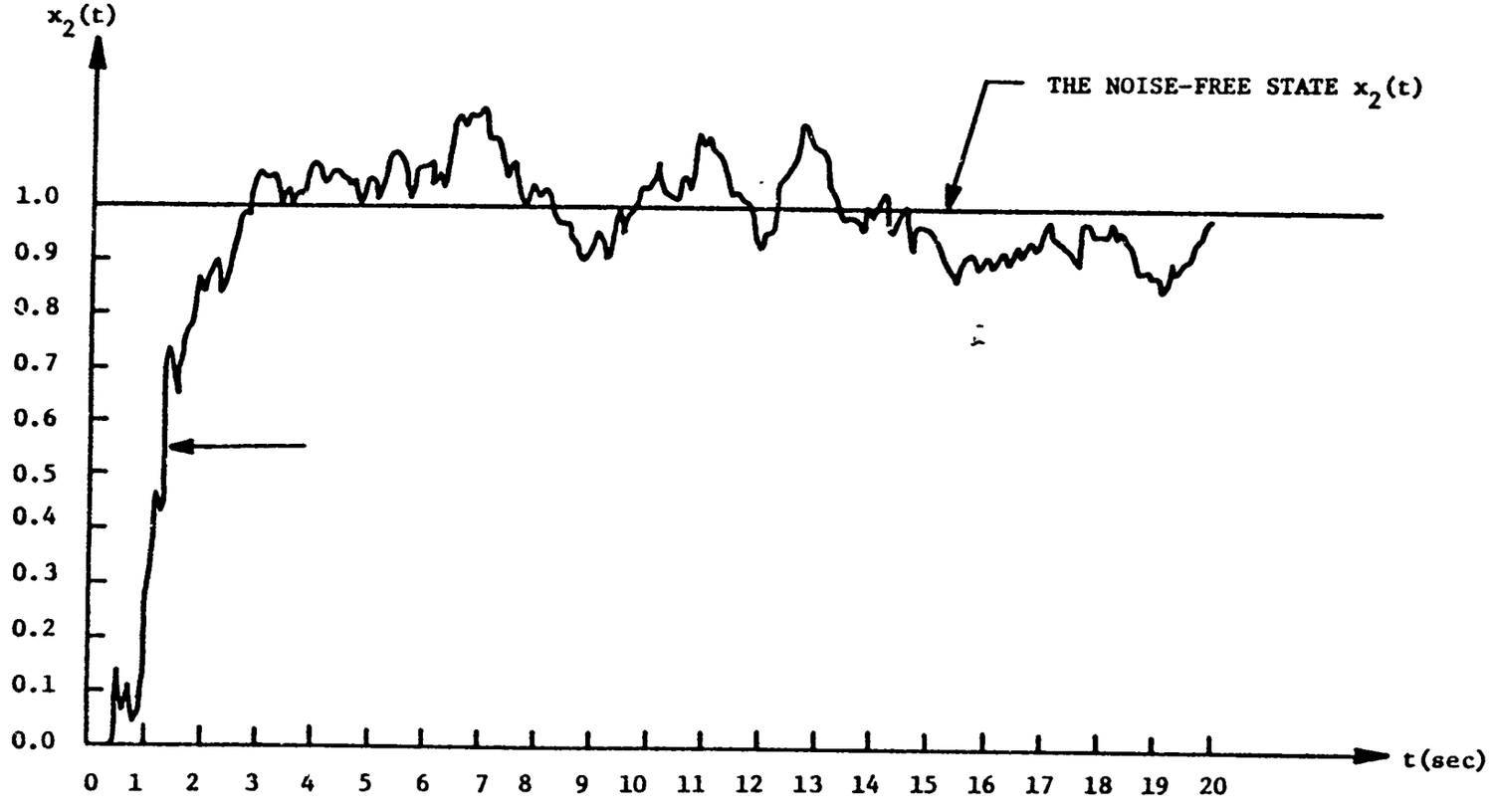


FIGURE 2. THE NOISE-FREE STATE $x_2(t)$ AND THE ESTIMATED STATE $\hat{x}_2(t)$

CHAPTER IV

APPLICATIONS OF PIECEWISE LINEAR APPROXIMATION

In Chapter III, piecewise constant approximation and piecewise linear approximation are used in finding the optimal gain and Kalman gain. In this chapter, we shall focus on the application of piecewise linear approximation.

Since the nineteenth century, the applications of piecewise constant functions, such as the block pulse function [26] and the Walsh function, [27] have been widely investigated by engineers in several fields, including optical engineering [28,29], biomedical sciences [30,31], communication theory [26,32,33], control system [34,35] and stochastic systems [36,37]. The advantages of these piecewise constant functions are that they introduce fairly accurate approximation techniques in analyzing electrical equipment and that they yield precise results in the simulation and design of a real system. The purpose of this chapter is to examine some important properties of piecewise linear functions and their extensions in system simulation.

From (2-2) the exact solution of a continuous-time state equation given in (2-1) is

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\lambda)}Bu(\lambda)d\lambda , \qquad (4-1)$$

where the second part in the right-hand side equation is the convolution integral between e^{At} and $Bu(t)$ for a causal system; therefore, we can rewrite (4-1) as follows:

$$x(t) = e^{At}x(0) + \int_0^t e^{A\lambda}Bu(t-\lambda)d\lambda \qquad (4\text{-}2)$$

Shieh [38] evaluated the states by using a rectangular approximation of the input, u(t), which is equivalent to inserting a sampler and a zero-order hold device before an integrator. The rectangular approximation, however, may be unsatisfactory when a stiff input is applied. In this case we may approximate the input by a series of piecewise linear functions.

If $t = KT$ and λ in (4-2) is in the range $[iT,(i+1)T]$ for $i = 0,1,2,\ldots,K-1$, then $u(t-\lambda)$ can be approximated in the following fashion:

$$\begin{aligned} u(t-\lambda) &= u(KT-\lambda) \\ &\simeq u[(K-i-1)T] + \frac{1}{T}\{u[(K-i)T]-u[(K-i-1)T]\}[KT-\lambda-(K-i-1)T] \\ &= u_{K-i-1} + \frac{1}{T}(u_{K-i}-u_{K-i-1})[(i+1)T-\lambda]. \end{aligned} \qquad (4\text{-}3)$$

Here we use u_j to represent $u(jT)$ for simplification of the derivation. Substituting (4-3) into (4-2), we find:

$$\begin{aligned} x(t) = x(KT) &= e^{At}x(0) + \int_0^t e^{A\lambda}Bu(t-\lambda)d\lambda \\ &\simeq e^{AKT}x(0) + \sum_{i=0}^{K-1}\int_{iT}^{(i+1)T} e^{A\lambda}Bu(t-\lambda)d\lambda \\ &\simeq e^{AKT}x(0) + \sum_{i=0}^{K-1}\int_{iT}^{(i+1)T} e^{A\lambda}B\{u_{K-1} + \frac{1}{T}(u_{K-i}-u_{K-i-1})[(i+1)T-\lambda]d\lambda \\ &= e^{AKT}x(0) + \sum_{i=0}^{K-1}\frac{1}{T}e^{AiT}\{[I-e^{AT}(I-AT)]A^{-2}Bu_{K-i-1} + [e^{AT}-(I+AT)]A^{-2}Bu_{K-} \end{aligned} \qquad (4\text{-}4)$$

A recursive relation is obtained from (4-4) by substituting t = (K+1)T, namely;

$$x[(K+1)T] = e^{A(K+1)T}x(0)+\sum_{i=0}^{K}\frac{1}{T}e^{AiT}\{[I-e^{AT}(I-AT)]A^{-2}Bu_{K-i}$$
$$+[e^{AT}-(I+AT)]A^{-2}Bu_{K-i+1}\}$$
$$= e^{AT}\{e^{AKT}x(0)+\sum_{i=1}^{K}\frac{1}{T}e^{A(i-1)T}[(I-e^{AT}(I-AT))A^{-2}Bu_{K-i}$$
$$+(e^{AT}-(I+AT))A^{-2}Bu_{K-i+1}]\}$$
$$+\frac{1}{T}[I-e^{AT}(I-AT)]A^{-2}Bu_{K}$$
$$+\frac{1}{T}[e^{AT}-(I+AT)]A^{-2}Bu_{K+1}$$
$$= e^{AT}x(KT)+M_{K}u(KT)+M_{K+1}u(K+1)\,. \qquad (4\text{-}5a)$$

where

$$M_{K} = \frac{1}{T}[I-e^{AT}(I-AT)]A^{-2}B \qquad (4\text{-}5b)$$

$$M_{K+1} = \frac{1}{T}[e^{AT}-(I+AT)]A^{-2}B. \qquad (4\text{-}5c)$$

However, by applying the piecewise constant approximation [38], the discrete state solution can be expressed as:

$$x[(K+1)T] = e^{AT}x(KT)+Mu^{*}(KT), \qquad (4\text{-}6a)$$

where

$$M = (e^{AT}-I)A^{-1}B \tag{4-6b}$$

and

$$u^*(KT) = \frac{1}{2}\{u(KT)+u[(K+1)]\}. \tag{4-6c}$$

By using G as given in (2-14), the corresponding M, M_K and M_{K+1} are formulated in Table 4-1.

If we choose the best approximation mode for e^{AT} given in (2-34), i.e.,

$$\begin{aligned} G &= e^{AT} \\ &= [(e^{\frac{-1}{2}AT_n})^{-1}(e^{\frac{1}{2}AT_n})]^n \\ &= \{(I-\frac{1}{2j}AT_n)[I+\sum_{i=1}^{j-1}\frac{(-1)^i\cdot(j-i)}{(2^i)\cdot(j)\cdot(i!)}(AT_n)^i]\}^{-n} * \\ &\quad \{(I+\frac{1}{2j}AT_n)[I+\sum_{i=1}^{j-1}\frac{(j-i)}{(2^i)\cdot(j)\cdot(i!)}(AT_n)^i]\}^n. \end{aligned} \tag{4-7}$$

By substituting (4-7) into (4-5b), (4-5c) and (4-6b), the results are:

$$\begin{aligned} M = T\{I+\sum_{i=1}^{j}\frac{(-1)^i[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)}(AT_n)^i\}^{-n} * \\ \{\sum_{i=1}^{j}\frac{[1-(-1)^i][j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)}(AT_n)^{i-1}\}^n B \end{aligned} \tag{4-8a}$$

n	G	M	M_K	M_{K+1}
1	$(I-AT)^{-1}$	$T(I-AT)^{-1}B$	$\frac{1}{2}T(I-AT)^{-1}B$	$\frac{1}{2}T(I-AT)^{-1}B$
2	$(I-\frac{1}{2}AT)^{-1}(I+\frac{1}{2}AT)$	$T(I-\frac{1}{2}AT)^{-1}B$	$\frac{1}{2}T(I-\frac{1}{2}AT)^{-1}B$	$\frac{1}{2}T(I-\frac{1}{2}AT)^{-1}B$
3	$(I-\frac{1}{3}AT)^{-1}[I+\frac{2}{3}AT$ $+\frac{1}{6}(AT)^2]$	$T(I-\frac{1}{3}AT)^{-1}\cdot$ $(I+\frac{1}{6}AT)B$	$\frac{1}{2}T(I-\frac{1}{3}AT)^{-1}\cdot$ $(I+\frac{1}{3}AT)B$	$\frac{1}{2}T(I-\frac{1}{3}AT)^{-1}B$
4	$(I-\frac{1}{4}AT)^{-1}[I+\frac{3}{4}AT$ $+\frac{1}{4}(AT)^2+\frac{1}{24}(AT)^3]$	$T(I-\frac{1}{4}AT)^{-1}[I+$ $\frac{1}{4}AT+\frac{1}{24}(AT)^2]B$	$\frac{1}{2}T(I-\frac{1}{4}AT)^{-1}[I+$ $\frac{5}{12}AT+\frac{1}{12}(AT)^2]B$	$\frac{1}{2}T(I-\frac{1}{4}AT)^{-1}\cdot$ $(I+\frac{1}{12}AT)B$
5	$(I-\frac{1}{5}AT)^{-1}[I+\frac{4}{5}AT$ $+\frac{3}{10}(AT)^2+\frac{1}{15}(AT)^3+\frac{1}{120}(AT)^4]$	$T(I-\frac{1}{5}AT)^{-1}[I+\frac{3}{10}AT$ $+\frac{1}{15}(AT)^2+\frac{1}{120}(AT)^3]B$	$\frac{1}{2}T(I-\frac{1}{5}AT)^{-1}[I+\frac{7}{15}AT$ $+\frac{7}{60}(AT)^2+\frac{1}{60}(AT)^3]B$	$\frac{1}{2}T(I-\frac{1}{5}AT)^{-1}[I+$ $\frac{2}{15}AT+\frac{1}{60}(AT)^2]B$

TABLE 4-1. THE DISCRETE-TIME SYSTEM MATRICES

$$M_K = T\{I+ \sum_{i=1}^{j} \frac{(-1)^i[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT_n)^i\}^{-n}*$$

$$\{-\sum_{i=2}^{j} \frac{[1-(-1)^i][j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT_n)^{i-2}+\sum_{i=1}^{j} \frac{[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT_n)^{i-1}\}B \quad (4\text{-}8b)$$

$$M_{K+1} = T\{I+ \sum_{i=1}^{j} \frac{(-1)^i\cdot[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT_n)^i\}^{-n}*$$

$$\{\sum_{i=2}^{j} \frac{[1-(-1)^i][j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT_n)^{i-2}-\sum_{i=1}^{j} \frac{(-1)^i[j^2-i(i-1)]}{(2^i)\cdot(j^2)\cdot(i!)} (AT_n)^{i-1}\}B \quad (4\text{-}8c)$$

The G, M, M_K and M_{K+1} matrices as obtained by substituting j = 1,2,3,4,5 in (4-7) and (4-8) are listed in Table 4-2.

In order to see the varied results obtained by using piecewise constant approximation (abbreviated PC) and piecewise linear approximation (abbreviated PL), we shall look into the following linearized 2-shaft gas turbine model developed by Mueller [39]:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} -1.268 & -0.04528 & 1.498 & 951.5 \\ 1.00197 & -1.957 & 8.52 & 1240 \\ 0 & 0 & -10 & 0 \\ 0 & 0 & 0 & -100 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 10 & 0 \\ 0 & 100 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad (4\text{-}9a)$$

n	G	M	M_K	M_{K+1}
1	$(I-\frac{1}{2}AT)^{-1}(I+\frac{1}{2}AT)$	$T(I-\frac{1}{2}AT)^{-1}B$	$\frac{1}{2}T(I-\frac{1}{2}AT)^{-1}B$	$\frac{1}{2}T(I-\frac{1}{2}AT)^{-1}B$
2	$(I-\frac{1}{2}AT+\frac{1}{16}(AT)^2]^{-1}\cdot$ $[I+\frac{1}{2}AT+\frac{1}{16}(AT)^2]$	$T[I-\frac{1}{2}AT+\frac{1}{16}(AT)^2]^{-1}B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{1}{16}(AT)^2]^{-1}$ $\cdot(I+\frac{1}{8}AT)B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{1}{16}(AT)^2]^{-1}$ $\cdot[I-\frac{1}{8}AT)B$
3	$[I-\frac{1}{2}AT+\frac{7}{72}(AT)^2$ $-\frac{1}{144}(AT)^3]^{-1}\cdot$ $[I+\frac{1}{2}AT+\frac{7}{72}(AT)^2$ $+\frac{1}{144}(AT)^3]$	$T[I-\frac{1}{2}AT+\frac{7}{72}(AT)^2$ $-\frac{1}{144}(AT)^3]^{-1}\cdot$ $[I+\frac{1}{72}(AT)^2B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{7}{72}(AT)^2$ $-\frac{1}{144}(AT)^3]^{-1}\cdot$ $[I+\frac{1}{6}AT+\frac{1}{72}(AT)^2B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{7}{72}(AT)^2$ $-\frac{1}{144}(AT)^3]^{-1}\cdot$ $[I-\frac{1}{6}AT+\frac{1}{72}(AT)^2]B$
4	$[I-\frac{1}{2}AT+\frac{7}{64}(AT)^2-\frac{5}{384}(AT)^3$ $+\frac{1}{1536}(AT)^4]^{-1}\cdot$ $[I+\frac{1}{2}AT+\frac{7}{64}(AT)^2+\frac{5}{384}(AT)^3$ $+\frac{1}{1536}(AT)^4]$	$T[I-\frac{1}{2}AT+\frac{7}{64}(AT)^2$ $-\frac{5}{384}(AT)^3+\frac{1}{1536}(AT)^4]^{-1}$ $\cdot[I+\frac{1}{192}(AT)^2]B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{7}{64}(AT)^2$ $-\frac{5}{384}(AT)^3+\frac{1}{1536}(AT)^4]^{-1}\cdot$ $[I+\frac{1}{6}AT+\frac{5}{192}(AT)^2$ $+\frac{1}{768}(AT)^3]B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{7}{64}(AT)^2$ $-\frac{5}{384}(AT)^3+\frac{1}{1536}(AT)^4]^{-1}\cdot$ $[I-\frac{1}{6}AT+\frac{1}{192}(AT)^2$ $-\frac{1}{768}(AT)^3]B$

TABLE 4-2. THE DISCRETE-TIME SYSTEM MATRICES [EQ. (2-34)]

n	G	M	M_K	M_{K+1}
5	$[I-\frac{1}{2}AT+\frac{23}{200}(AT)^2-\frac{19}{1200}(AT)^3$ $+\frac{13}{9600}(AT)^4-\frac{1}{19200}(AT)^5]^{-1}\cdot$ $[I+\frac{1}{2}AT+\frac{23}{200}(AT)^2+\frac{19}{1200}(AT)^3$ $+\frac{13}{9600}(AT)^4+\frac{1}{19200}(AT)^5]$	$T[I-\frac{1}{2}AT+\frac{23}{200}(AT)^2-\frac{19}{1200}(AT)^3$ $+\frac{13}{9600}(AT)^4-\frac{1}{19200}(AT)^5]^{-1}\cdot$ $[I+\frac{19}{600}(AT)^2+\frac{1}{9600}(AT)^4]B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{23}{200}(AT)^2-\frac{19}{1200}(AT)^3$ $+\frac{13}{9600}(AT)^4-\frac{1}{19200}(AT)^5]^{-1}\cdot$ $[I+\frac{1}{6}AT+\frac{19}{600}(AT)^2$ $+\frac{1}{9600}(AT)^4]B$	$\frac{1}{2}T[I-\frac{1}{2}AT+\frac{23}{200}(AT)^2$ $-\frac{19}{1200}(AT)^3+\frac{13}{9600}\cdot$ $(AT)^4-\frac{1}{19200}(AT)^5]^{-1}\cdot$ $[I-\frac{1}{6}AT+\frac{19}{600}(AT)^2$ $+\frac{1}{9600}(AT)^4]B$

TABLE 4-2. THE DISCRETE-TIME SYSTEM MATRICES [EQ. (2-34)] CONTINUED

where

$$u_1(t) = \text{unit step input} \tag{4-9b}$$

$$u_2(t) = \text{ramp input}. \tag{4-9c}$$

The initial conditions are:

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \\ x_4(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{4-9d}$$

The exact state solutions can be found as:

$$x_1(t) = -541.752+714.701t+549.442e^{-1.34174t} - 7.771e^{-1.88326t}+0.177e^{-10t}-0.096e^{-100t} \tag{4-10a}$$

$$x_2(t) = -790.109+999.545t+894.784e^{-1.34174t} - 105.586e^{-1.88326t}+1.037e^{-10t}-0.125e^{-100t} \tag{4-10b}$$

$$x_3(t) = 1-e^{-10t} \tag{4-10c}$$

$$x_4(t) = -0.01+t+0.0008e^{-1.34174t}-0.0004e^{-1.88326t} + 0.01e^{-100t}. \tag{4-10d}$$

For the sake of comparison , we choose the approximated state transition matrix as (n=2 in Table 4-2) :

$$G = [I - \frac{1}{2} AT + \frac{1}{16} (AT)^2]^{-1}[I + \frac{1}{2} AT + \frac{1}{16} (AT)^2] . \qquad (4\text{-}11)$$

The states obtained by using PC and PL approximations are listed in Tables 4-3, 4-4, 4-5 and 4-6. Comparing these values with the exact state values, we conclude that PL gives a better result than PC. Therefore, if the polygonal hold, (a device for integration using the trapezoidal approximation method [40]) can be realized, a more accurate discrete-time state solution can be achieved.

Time	Exact x_1	x_1 (Using PC)	x_1 (Using PL)
0	0	0	0
0.02	0.08446	0.10736	0.08657
0.04	0.47723	0.49992	0.47736
0.06	1.22978	1.25147	1.22925
0.08	2.34013	2.36100	2.33933
0.10	3.79942	3.81955	3.79844
0.12	5.59816	5.61758	5.59702
0.14	7.72708	7.74581	7.72580
0.16	10.1772	10.1953	10.1758
0.18	12.9398	12.9573	12.9383
0.20	16.0066	16.0234	16.0049

TABLE 4-3. COMPARISONS OF APPROXIMATED x_1 (BY USING PC AND PL) WITH THE EXACT x_1

Time	Exact x_2	x_2 (Using PC)	x_2 (Using PL)
0	0	0	0
0.02	0.12230	0.15211	0.12496
0.04	0.66790	0.69748	0.66796
0.06	1.70049	1.72882	1.69966
0.08	3.21492	3.24226	3.21375
0.10	5.19778	5.22422	5.19637
0.12	7.63529	7.66088	7.63370
0.14	10.5144	10.5392	10.5126
0.16	13.8227	13.8467	13.8207
0.18	17.5484	17.5716	17.5463
0.20	21.6804	21.7029	21.6782

TABLE 4-4. COMPARISONS OF APPROXIMATED x_2 (BY USING PC AND PL) WITH THE EXACT x_2

Time	Exact x_3	x_3 (Using PC)	x_3 (Using PL)
0	0	0	0
0.02	0.18127	0.18141	0.18141
0.04	0.32968	0.32990	0.32990
0.06	0.45119	0.45146	0.45146
0.08	0.55067	0.55097	0.55097
0.10	0.63212	0.63243	0.63243
0.12	0.69881	0.69911	0.69911
0.14	0.75340	0.75369	0.75369
0.16	0.79810	0.79837	0.79837
0.18	0.83470	0.83495	0.83495
0.20	0.86467	0.86489	0.86489

TABLE 4-5. COMPARISONS OF APPROXIMATED x_3 (BY USING PC AND PL) WITH THE EXACT x_3

Time	Exact x_4	x_4 (Using PC)	x_4 (Using PL)
0	0	0	0
0.02	0.01135	0.00889	0.01111
0.04	0.03018	0.02765	0.03012
0.06	0.05	0.04752	0.05001
0.08	0.07	0.06750	0.07
0.10	0.09	0.08750	0.09
0.12	0.11	0.10750	0.11
0.14	0.13	0.12750	0.13
0.16	0.15	0.14750	0.15
0.18	0.17	0.16750	0.17
0.20	0.19	0.18750	0.19

TABLE 4-6. COMPARISONS OF APPROXIMATED x_4 (BY USING PC AND PL) WITH THE EXACT x_4

CHAPTER V

CONCLUSIONS

A simple and fast algorithm is developed for approximate linear regulator and Kalman filter problems.

After introducing the definitions for the linear regulator and the Kalman filter, we establish piecewise-constant and piecewise linear approximations to solve for the state transition matrix. A geometrical series approach with scaling and squaring is used in approximating the exponential of a matrix. Piecewise-constant gains and piecewise time-varying gains for approximating a time-varying optimal linear gain and a time-varying Kalman gain of quadratic synthesis problems can be solved through this approach.

The proposed method greatly improves the accuracy and computational speed of the existing methods which use the Walsh function and the block-pulse function. The developed suboptimal feedback gains for a deterministic continuous system and the approximate Kalman gains for a continuous stochastic system can be readily implemented on the low-cost microprocessors or minicomputers for on-line control and estimation.

The effectiveness of the piecewise-linear approximation is further demonstrated by examining the system simulation problem. A transformation from a continuous-time system equation to a discrete-time system equation is derived using a piecewise-linear approximation technique. The result is surprisingly accurate. The errors of simulating a linearized 2-shaft gas turbine model with only 11 samples are within 0.03 percent.

The Linear regulator problem and the Kalman filter problem are two of

the most often encountered control problems. The proposed algorithm has been found to be efficient in solving these problems. Potential usefulness of this method in solving other problems remains to be exploited.

BIBLIOGRAPHY

[1] M. Athans and P. L. Falb, Optimal Control, McGraw-Hill, New York, 1966.

[2] R. A. Miller, Specific optimal control of the linear regulator using a minimal order observal, Int. J. Control, 18, July 1973, pp. 139-159.

[3] N. Wiener, The Extrapolation, Interpolation and Smoothing of Stationary Time Series. John Wiley & Sons, New York, 1948.

[4] R. E. Kalman, A new approach to linear filtering and prediction problems, ASME J. Basic Eng. 82, 1960, pp. 34-45.

[5] R. E. Kalman and R. S. Bucy, New results in linear filtering and prediction theory, ASME J. Basic Eng. 83, 1961, pp. 95-107.

[6] K. J. Åström, Introduction to Stochastic Control Theory, Academic Press, 1970.

[7] T. Kailath, An innovation approach to least-square estimation part I: linear filtering in additive white noise, IEEE Trans. Autom. Control AC-13, 1968, pp. 646-655.

[8] T. Kailath and P. Frost, An innovation approach to least-square estimation part II: linear smoothing in additive white noise, IEEE Trans. Autom. Control AC-13, 1968, pp. 655-660.

[9] A. E. Bryson and Y. C. Ho, Optimal Programming, Estimation and Control, Blaisdell, New York, 1968.

[10] B. C. Kuo, Analysis and Synthesis of Sampled-Data Control Sysrem, Prentice-Hall, New Jersey, 1963.

[11] G. E. F. Sherwood and A. E. Taylor, Calculus, Prentice-Hall, New York, 1952, pp. 371-392.

[12] L. S. Shieh, R. E. Yates, and J. M. Navarro, Solving inverse Laplace transform, linear and nonlinear state equations using block-pulse functions, Computer and Electrical Engineering, Pergamon Press, Vol. 6, No. 1, March 1979, pp. 3-18.

[13] M. Vidyasagar, Nonlinear System Analysis, Prentice Hall, 1978, pp. 113-117.

[14] C. B. Moler and C. F. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, SIAM Review, Vol. 20, No. 4, October 1978, pp. 801-836.

[15] A. Gelb, J. F. Kasper, R. A. Nash, C. F. Price and A. A. Sutherland, Applied Optimal Estimation, The MIT Press, Cambridge, Massachusetts, 1974.

[16] D. L. Kleinman, T. Fortman and M. Athans, On the design of linear systems with piecewise constant feedback gains, IEEE Trans. on Automatic Control, Vol. AC-13, August 1968, pp. 354-361.

[17] C. F. Chen and C. H. Hsiao, Design of piecewise constant gains for optimal control via Walsh functions, IEEE Trans. on Automatic Control, Vol. AC-20, October 1975, pp. 596-603.

[18] V. P. Rao and K. R. Rao, Optimal feedback control via block-pulse functions, IEEE Trans. on Automatic Control, Vol. AC-24, No. 2, April 1979, pp. 372-374.

[19] A. E. Bryson and Y. C. Ho, Applied Optimal Control, Ginn and Company, Waltham, Massachusetts, 1969.

[20] D. G. Lainiotis, Partitioned Riccati solutions and integration-free doubling algorithms, IEEE Trans. on Automatic Control, Vol. AC-21, Oct. 1976, pp. 677-687.

[21] Y. Takahashi, M. Tomizuka and D. M. Auslander, Simple discrete control of industrial processes, Trans. on ASME J. of Dynamic Systems, Measurement and Control, Vol. 97, No. 4, December 1975, pp. 354-361.

[22] D. M. Auslander, Y. Takahashi and M. Tomizuka, Direct digital process control: practice and algorithm for microprocessor application, Proc. IEEE, Vol. 66, No. 2, February 1978, pp. 199-208.

[23] E. B. Dahlin, Designing and tuning digital controllers, Instrument & Control Systems, Vol. 42, June 1968, pp. 77-83.

[24] L. S. Shieh, C. K. Young and B. C. McInnis, Solutions of state space equations via block-pulse functions, Int. J. of Control, Vol. 28, No. 3, 1978, pp. 382-392.

[25] C. F. Chen and C. H. Hsiao, A state space approach to Walsh series solutions of linear systems, Int. J. System Science, Vol. 6, No. 9, 1975, pp. 833-858.

[26] H. F. Harmuth, Application of Walsh functions in communication, IEEE Spectrum, 1969, pp. 82-91.

[27] J. L. Walsh, A closed set of orthogonal functions, Ann. Journ. Math., Vol. 55, 1923, pp. 5-24.

[28] J. D. Kennedy, Walsh function imagery analysis, Proc. Walsh Function Symp., Nav. Res. Labs., Washington, D. C., 1971, pp. 7-10.

[29] P. J. Milne, N. Ahmed, R. R. Gallagher and S. G. Harris, An application of Walsh function to the monitoring of electrocardiograph signals, Proc. Walsh Function Symp., Nav. Res. Labs., Washington, D. C., 1972, pp. 149-153.

[30] C. W. Thomas and A. J. Welch, Heart rate representation using Walsh functions, Proc. Walsh Function Symp., Nav. Res. Labs., Washington, D. C., 1972, pp. 154-158.

[31] H. C. Andrews, Walsh function from the perspective of useful unitary transformations for data processing, Proc. of IEEE Conference on Decision and Control, 1972, pp. 492-494.

[32] J. D. Lee, Review of recent work on applications of Walsh functions in communications, Proc. Walsh Function Symp., Nav. Res. Labs., Washington, D. C., 1970, pp. 26-35.

[33] J. E. Gibbs and H. A. Gebbie, Application of Walsh function to transform spectroscopy, Nature, Vol. 224, Dec. 1969, pp. 1012-1013.

[34] F. Picher, Walsh function and linear system theory, Proc. Walsh Function Symp., Nav. Res. Labs., Washington, D. C., 1970, pp. 175-182.

[35] J. E. Gibbs and M. J. Millard, Some methods of solution of linear ordinary logical differential equations, DES Report, No. 2, Nat. Phys. Lab., Teddington, Middlesex, England, 1969.

[36] M. Maqusi, On moments and Walsh characteristic functions, IEEE Trans. on Communications, Vol. COM-21, June 1973, pp. 768-770.

[37] A. S. French and E. G. Butz, The use of Walsh function in the Wiener analysis of nonlinear systems, IEEE Trans. on Computers, Vol. C-23, No. 3, March 1974, pp. 225-232.

[38] L. S. Shieh, R. E. Yates and J. M. Navarro, Representation of continuous-time state equations by discrete-time state equations, IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-8, No. 6, June 1978, pp. 485-492.

[39] G. S. Mueller, Linear model of a two shaft turbo jet and its properties, Proc. IEE, Vol. 117, No. 10, 1970, pp. 2050-2056.

[40] E. I. Jury, Theory and Application of the Z-Transform Method, John Wiley & Sons, New York, 1964.

[41] L. S. Shieh, W. B. Wai and R. E. Yates, Approximate Kalman filters, JACC 1980 (to be presented).

[42] L. S. Shieh, W. B. Wai and R. E. Yates, A geometric series approach for approximation of transition matrices in quadratic synthesis, ASME 1980 (to be published).

Determination of Equivalent Dominant Poles and Zeros Using Industrial Specifications

LEANG-SAN SHIEH, MEMBER, IEEE, YING-JYI PAUL WEI, MEMBER, IEEE, HSI-ZEN CHOW, AND ROBERT E. YATES

Abstract—A graphical method and an analytical method are presented to determine the equivalent dominant poles and zeros of a system using assigned industrial specifications. A second-order transfer function with two poles and one finite zero is used to investigate the relationships between industrial specifications and the two poles and one finite zero. Also, it is used to verify the rule of the thumb obtained from Axelby's empirical results. A frequency response data matching method is proposed for fitting a low-order transfer function using the assigned industrial specifications that are obtained from a given high-order transfer function. Thus the equivalent dominant poles and zeros of a high-order system can be determined from the identified low-order model.

I. Introduction

In the filter and compensator designs it is necessary and useful to have a rapid method or a simple graphical method to determine the poles and zeros that dominate the characteristics of the transient response. These poles and zeros are called the dominant poles and zeros that can be used to estimate the dynamic behavior of the system response. In the literature, the definitions of the dominant poles and zeros are ambiguous. For example, the dominant poles are commonly defined as the poles which are located near the imaginary axis (the $j\omega$ axis) or the poles which have the smallest absolute value when no significant zeros appear. Sometimes a pole P_1 is defined as the dominant pole [1] if $|P_i| \geq 5|P_1|$ where P_i are other system poles. The roles of dominant zeros that are often neglected in the literature become significant if the precise dynamic characteristics of a system in the transient state are required. The zeros not only contribute to the initial conditions of the transient response but also increase the bandwidth in the frequency domain; therefore, the roles of the zeros are as important as those of the poles.

As the technologies are progressing, the accurate description of many physical systems results in a high-order transfer function that consists of many clustery poles and zeros in the s plane. The poles near the $j\omega$ axis may not be dominant poles because the dominant effects on the transient response behavior of the poles are cancelled by the nearby zeros, and the system response may be characterized by the collective efforts of a group of clustery poles and zeros. This implies that the poles and zeros which are not near the $j\omega$ axis may dominate the characteristics of the system response. Therefore, the equivalent dominant poles and zeros, rather than the dominant poles and zeros obtained from the geometric locations in the s plane, become significant in the analysis and synthesis of a high-order system. Furthermore, the design goals and the nature of a high-order system are often characterized by a set of control specifications [2] (called the industrial specifications) that are commonly classified as 1) the time-domain specifications, for example, the rise time and the overshoot, 2) the frequency-domain specifications, for example, the bandwidth and the phase margin, 3) the complex-domain specifications, for example, the damping ratio and the undamped natural angular frequency or the equivalent poles and zeros in the s plane. If the relationships among the time-domain, frequency-domain specifications, and the equivalent poles and zeros (the complex-domain specifications) can be simply determined from a simple equation or a working graph, then the selected poles and zeros in the design of filters and compensators become meaningful, and the design processes can be greatly simplified.

In this paper, a graphical method and an analytical method are proposed to determine the equivalent dominant poles and zeros using assigned industrial specifications. First, relationships among various industrial specifications will be studied. A second-order transfer function having two poles and one finite zero is used as a basis for the investigation. Several working graphs and mathematical expressions are developed for the determination of the two dominant poles and one dominant zero using the assigned industrial specifications. Then the equivalent dominant poles and zeros of a high-order system are determined by a new dominant frequency-response data matching method. The equivalent dominant poles and zeros thus obtained satisfy the exact assigned industrial specifications.

II. The Relationships Among Various Industrial Specifications

In control system design, the design goals are usually expressed in terms of a set of industrial specifications. The placement of poles and zeros based upon the assigned specifications needs certain experiences. If the relationships among various industrial specifications can be determined, then nonconflicting industrial specifications can be assigned as design goals, and the meaningful dominant poles and zeros can be selected for filter and compensator designs. Thus an effective design method may be developed.

Manuscript received July 13, 1978; revised January 25, 1979. This work was supported in part by U.S. Army Missile Command, Redstone Arsenal, AL, DAAK 00-79-C-0061, and U.S. Army Research Office DAAG29-77-G-0143.

L. S. Shieh, Y. J. Wei, and H. Z. Chow are with the Department of Electrical Engineering, University of Houston, Houston, TX 77004

R. E. Yates is with the Guidance and Control Directorate, U.S. Army Missile Research and Development Command, Redstone Arsenal, AL 35809.

0018-9421/79/0800-0125$00.75

An empirical study on the relationships among various industrial specifications has been conducted by Axelby [3]. The empirical rules or the rule of the thumb, which link the specifications in both time and frequency domains, are listed as follows:

$$M_t \simeq M_p \simeq \frac{1}{\sin \phi_m} \tag{1a}$$

where M_t is the maximum value of unit-step response, M_p is the maximum value of the closed-loop frequency response, and ϕ_m is the phase margin;

$$M_e \simeq \frac{1}{\omega_c} \tag{1b}$$

where M_e is the maximum value of the error of the unit-ramp function and ω_c is the gain crossover frequency;

$$\omega_p \simeq \omega_c \tag{1c}$$

where ω_p is the peak value frequency or the frequency when M_p occurs;

$$\dot{M}_t \simeq \omega_c \tag{1d}$$

where $\dot{M}_t$ is the maximum value of the unit-impulse response;

$$t_p \simeq \frac{3}{\omega_c} \tag{1e}$$

where t_p is the peak value time or the time when M_t occurs,

$$t_v \simeq \frac{1.8}{\omega_c} \tag{1f}$$

where t_v is the time when the maximum error of the ramp function with respect to its input occurs;

$$t_c \simeq \frac{1}{\omega_c} \tag{1g}$$

where t_c is the time when $\dot{M}_t$ occurs.

Other rules of the thumb according to Truxal [4] are listed as follows:

$$t_r \omega_b \simeq 0.6\pi \text{ to } 0.9\pi \tag{1h}$$

where t_r is the rise time or the time required for the response to go from 10 to 90 percent of its final value and ω_b is the bandwidth in rad/s;

$$t_d \simeq \frac{1}{K_v} \tag{1i}$$

where t_d is the delay time or the time required to reach 50 percent of its final value and K_v is the velocity error constant.

Some other analytical results that represent the relationships between the time-domain specifications (but not the frequency-domain specifications) and the complex-domain specifications have been developed and can be found in standard textbooks [5], [6]. The most commonly used function for investigating the relationships is

$$\frac{Y(s)}{R(s)} = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} \tag{2}$$

where $Y(s)$ and $R(s)$ are the output and input functions, respectively, and ξ is the damping ratio and ω_n is the undamped natural angular frequency. From (2) we observe that the zero of the system is located at infinity, and is not a significant zero. Since the time-domain specifications are often used to define the characteristics of the transient behavior, the roles of zeros become significant. Therefore, a better model than that of (2) should be used to study the relationships among the industrial specifications. The transfer function of a unit-feedback system that has two poles and one finite zero is used as a basis for the investigation. The proposed closed-loop transfer function is then,

$$\frac{Y(s)}{R(s)} = T(s) = \frac{\tau\omega_n s + \omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} = \frac{b_1 s + b_2}{s^2 + a_1 s + a_2} = \frac{B(s)}{A(s)}$$

$$= \frac{\tau\left(\frac{s}{\omega_n}\right) + 1}{\left(\frac{s}{\omega_n}\right)^2 + 2\xi\left(\frac{s}{\omega_n}\right) + 1} = \frac{\tau s^* + 1}{(s^*)^2 + 2\xi s^* + 1} \tag{3}$$

where s^* is a normalized complex variable, a_i and b_i are constants, and $A(s)$ and $B(s)$ are two polynomials. The normalized poles and the original poles are at

$$s_1^* = -\xi + j\sqrt{1-\xi^2} \qquad s_1 = -\xi\omega_n + j\omega_n\sqrt{1-\xi^2}$$
$$s_2^* = -\xi - j\sqrt{1-\xi^2} \qquad s_2 = -\xi\omega_n - j\omega_n\sqrt{1-\xi^2} \tag{4a}$$

and the normalized zero and the original zero are at

$$s^* = \frac{1}{\tau} \qquad s = -\frac{\omega_n}{\tau} \tag{4b}$$

The open-loop transfer function $G(s)$ of the system in (3) is

$$G(s) = \frac{\tau\omega_n s + \omega_n^2}{s[s + (2\xi\omega_n - \tau\omega_n)]} = \frac{K_v\left(1 + \frac{s}{b}\right)}{s\left(1 + \frac{s}{a}\right)} \tag{5}$$

where $K_v = \omega_n/(2\xi - \tau)$ is the velocity error constant if $\tau < 2\xi$

$a = (2\xi - \tau)\omega_n$ and $b = \omega_n/\tau$.

Comparing (2) and (3) we observe that a finite zero has been inserted in (3). The zero contributes the initial condition at the transient state, and it reduces the velocity error at the steady state. Also it provides an additional bandwidth in the frequency domain, which increases the phase margin and improves the stability of a system.

The derivations of the relationships among the industrial specifications are shown as the following seven relationships.

1) The Relationships Among M_t, t_p, ξ, ω_n, and τ The unit-step response of the system in (3) gives

$$Y(s) = \frac{\tau\omega_n s + \omega_n^2}{s(s^2 + 2\xi\omega_n s + \omega_n^2)} \tag{6a}$$

The inverse Laplace transform of $Y(s)$ results in

$$Y(t) = 1 - e^{-\xi\omega_n t}\left[\cos\omega_n\sqrt{1-\xi^2}\,t + \frac{\xi - \frac{\tau}{\omega_n}}{\sqrt{1-\xi^2}}\sin\omega_n\sqrt{1-\xi^2}\,t\right]. \tag{6b}$$

Differentiating $y(t)$ with respect to t and setting the result equal to zero yields

$$t_p = \left(\pi + \tan^{-1}\frac{\tau\sqrt{1-\xi^2}}{\tau\xi - 1}\right) \Big/ \left(\omega_n\sqrt{1-\xi^2}\right). \tag{6c}$$

Substituting (6c) into (6b) and simplifying it gives the maximum value of the unit-step response

$$M_t = 1 + e^{-\xi\omega_n t_p}(\tau^2 - 2\tau\xi + 1)^{1/2}. \tag{6d}$$

2) The Relationships Among M_p, ω_p, ξ, ω_n, and τ: Applying Higgins and Siegel's complex variable differentiation method [7], we can solve the peak value frequency ω_p from the following equation:

$$R_e\left\{j\left[\frac{1}{B(s)}\frac{dB(s)}{ds} - \frac{1}{A(s)}\frac{dA(s)}{ds}\right]\right\}_{s=j\omega_p} = 0. \tag{7a}$$

Thus we have

$$\left.\begin{aligned}\omega_p &= \omega_n\sqrt{1-2\xi^2}\\ M_p &= 1/(2\xi\sqrt{1-\xi^2})\end{aligned}\right\}, \quad \text{if } \tau = 0 \tag{7b}$$

and

$$\left.\begin{aligned}\omega_p &= \frac{\omega_n}{\tau}\left[-1 + \sqrt{(\tau^2+1)^2 - 4\tau^2\xi^2}\right]^{1/2}\\ M_p &= \frac{\tau^2}{\sqrt{2}}\left[\sqrt{(\tau^2+1)^2 - 4\xi^2\tau^2} - (\tau^2+1) + 2\xi^2\tau^2\right]^{-1/2}.\end{aligned}\right\}, \quad \text{if } \tau \neq 0. \tag{7c}$$

3) The Relationships Among ϕ_m, ω_c, ξ, ω_n, and τ: Using the definitions of ϕ_m and ω_c,

$$\phi_m = \angle G(s)\big|_{s=j\omega_c} + 180^\circ \tag{8a}$$

and

$$|G(s)|_{s=j\omega_c} = 1 \tag{8b}$$

we have

$$\phi_m = \tan^{-1}\left[\frac{\left(\frac{\omega_c}{\omega_n}\right)\tau + (2\xi - \tau)\left(\frac{\omega_n}{\omega_c}\right)}{1 - (2\xi - \tau)\tau}\right] \tag{8c}$$

and

$$\omega_c = \omega_n\left[2\xi\tau - 2\xi^2 + \sqrt{(2\xi^2 - 2\xi\tau)^2 + 1}\right]^{1/2}. \tag{8d}$$

4) The Relationships Among t_v, M_e, ξ, ω_n, and τ: The error signal $e(t)$, which is the difference between the ramp input $r(t)$ and the time response $y(t)$ of the same input to the system in (3), is

$$e(t) = \frac{2\xi - \tau}{\omega_n} - \frac{1}{A\omega_n}e^{-\xi\omega_n t}\left[B\cos\omega_n\sqrt{1-\xi^2}\,t - C\sin\omega_n\sqrt{1-\xi^2}\,t\right] \tag{9a}$$

where

$$A = (1-\xi^2),\ B = (2\xi - \tau)(1-\xi^2),$$
$$C = (1 - 2\xi^2 + \tau\xi)\sqrt{1-\xi^2}.$$

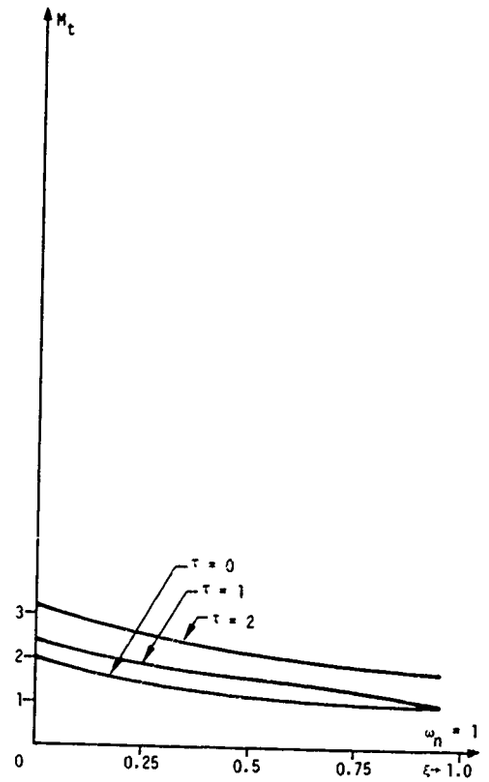


Fig. 1. Relationships among M_t, ξ, ω_n, and τ shown in (6d).

Differentiating $e(t)$ with respect to t and setting the result equal to zero we have

$$t_v = \frac{1}{\omega_n\sqrt{1-\xi^2}}\tan^{-1}\left[\frac{\sqrt{1-\xi^2}}{\tau - \xi}\right]. \tag{9b}$$

Substituting the t_v into (9a) and simplifying it we have

$$M_e = [2\xi - \tau + \sqrt{(1 + \tau^2 - 2\tau\xi)}e^{-\xi\omega_n t_v}]/\omega_n. \tag{9c}$$

5) The Relationships Among t_c, $\dot{M}_t$, ξ, ω_n, and τ: Differentiating the unit-impulse response $y(t)$ of the system in (3), $\dot{y}(t)$, and setting the result equal to zero, we have the time t_c at which the maximum value occurs, or

$$t_c = \frac{1}{\omega_n\sqrt{1-\xi^2}}\tan^{-1}\left[\frac{(1-2\xi\tau)\sqrt{1-\xi^2}}{\xi - 2\tau\xi^2 + \tau}\right]. \tag{10a}$$

Substituting t_c into $\dot{y}(t)$ yields the maximum value of the unit-impulse response $\dot{M}_t$, or

$$\dot{M}_t = \omega_n e^{-\xi\omega_n t_c}\sqrt{\tau^2 - 2\xi\tau + 1}. \tag{10b}$$

6) The Relationships Among K_v, ξ, ω_n, and τ: The velocity error constant K_v can be derived from the basic definition as

$$K_v = \lim_{s\to 0} s\cdot G(s) = \frac{\omega_n}{2\xi - \tau}, \quad \text{if } \tau < 2\xi. \tag{11}$$

7) The Relationships Among ω_b, ξ, ω_n, and τ: The definition of the bandwidth of a system is

$$|T(s)|_{s=j\omega_b} = \frac{1}{\sqrt{2}}. \tag{12}$$

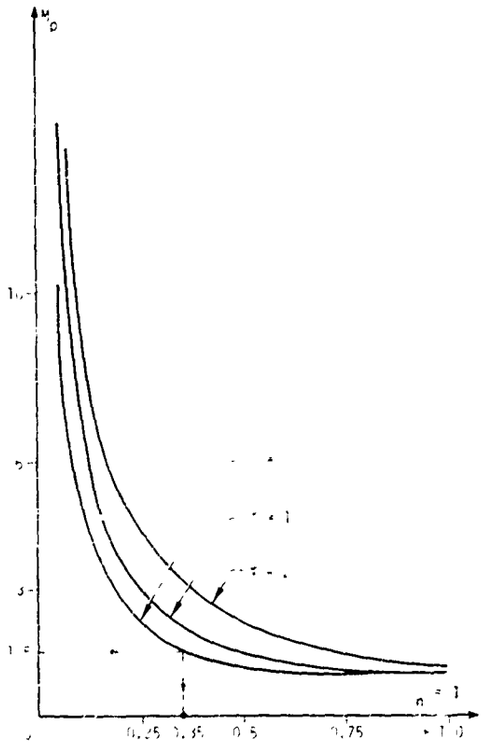


Fig. 2. Relationships among M_p, ξ, ω_n, and τ shown in (7).

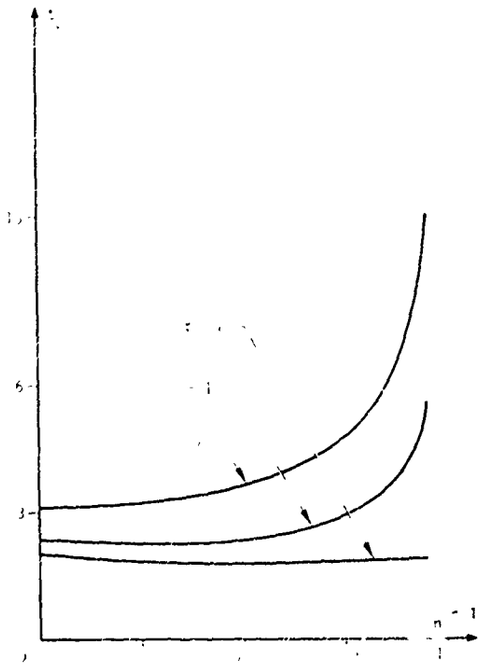


Fig. 3. Relationships among t_p, ξ, ω_n, and τ shown in (6c)

The analytical expression is

$$\omega_b = \omega_n [(1 + \tau^2 \quad 2\xi^2) + \sqrt{(1 + \tau^2 \quad 2\xi^2)^2 + 1}]^{1/2}. \quad (13)$$

Most important time-domain and frequency-domain specifications have been analytically expressed in terms of ξ, ω_n, and

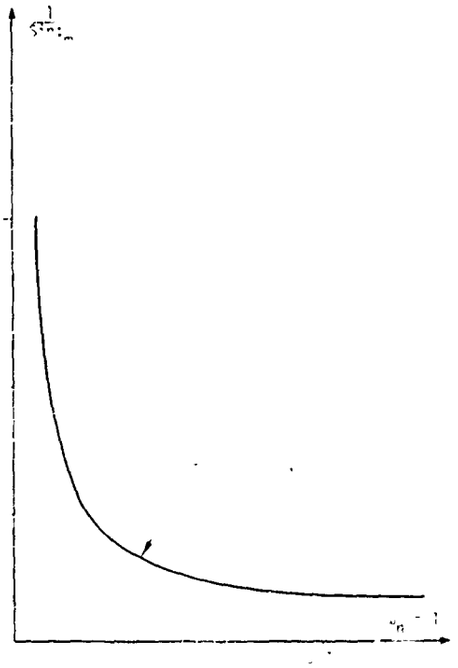


Fig. 4 Relationships among $1/\sin\phi_m$, ξ, ω_n, and τ shown in (8c)

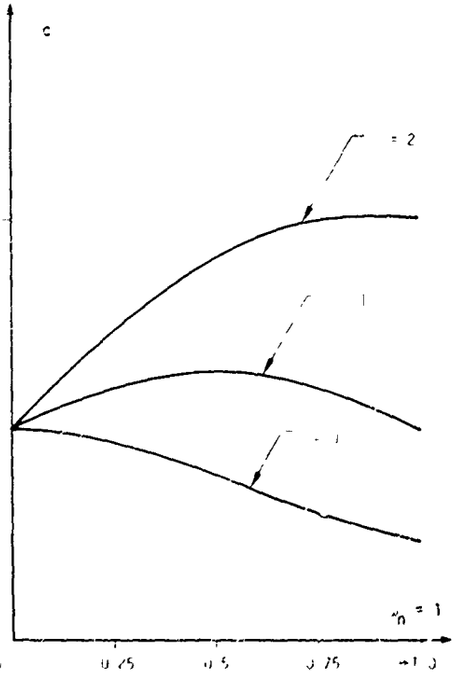


Fig. 5 Relationships among ω_c, ξ, ω_n, and τ shown in (8d).

τ which are the specifications in the complex domain. These expressions are normalized and graphically shown in Figs. 1-11. If an industrial specification is assigned, the corresponding ξ and τ or the equivalent poles and zero in (4) can be determined from the plotted curves. Also the curves in Figs. 12-15 can be used to verify the rules of the thumb proposed by Axelby [3]. It is observed that the accuracy of the rules de-

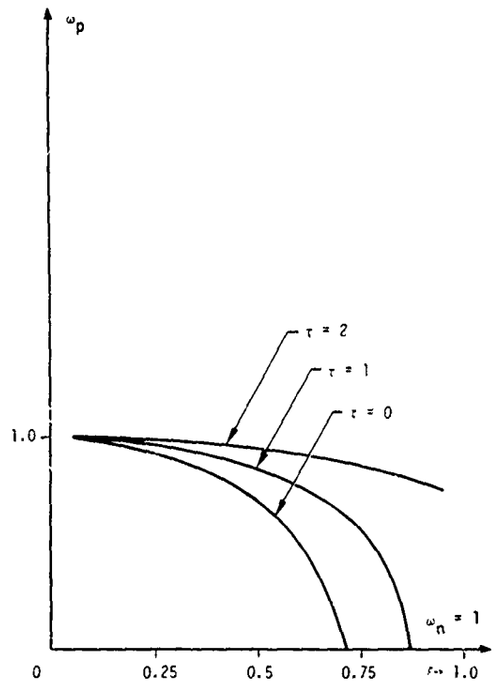


Fig. 6. Relationships among ω_p, ξ, ω_n, and τ shown in (7).

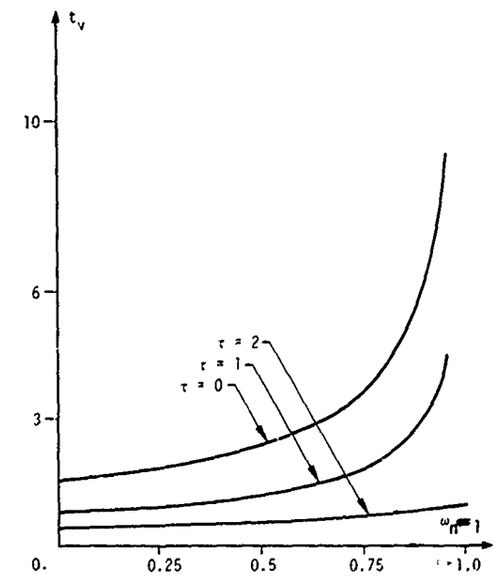


Fig. 7. Relationships among t_v, ξ, ω_n, and τ shown in (9b).

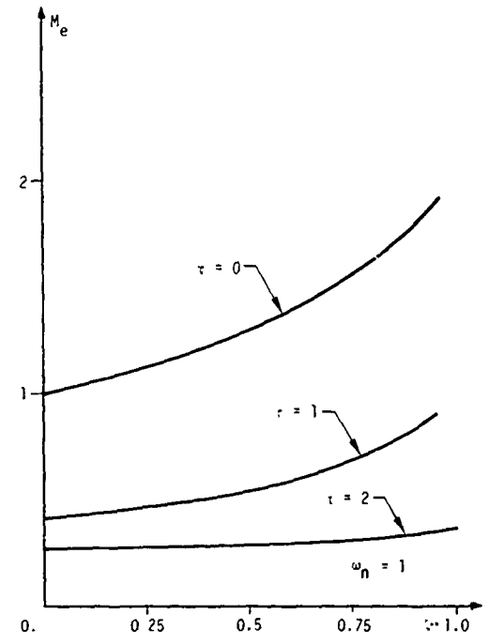


Fig. 8. Relationships among M_e, ξ, ω_n, and τ shown in (9c).

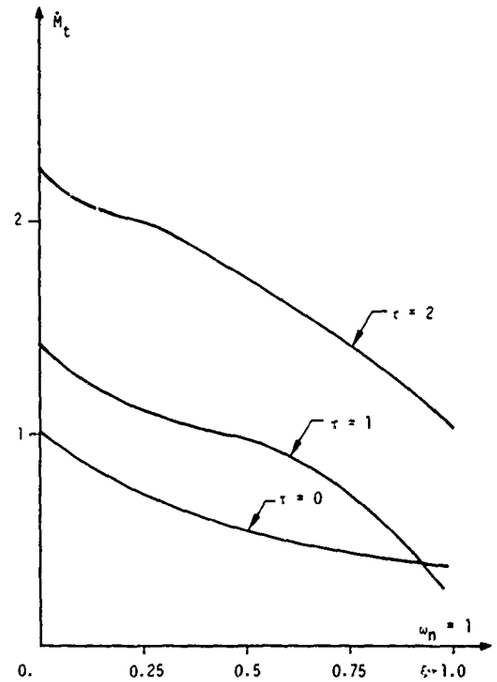


Fig. 9. Relationships among $\dot{M}_t$, ξ, ω_n, and τ shown in (10b).

pends upon the range of the damping ratio and the zero location. Furthermore, from the developed working graphs, a set of meaningful and nonconflicting specifications can be assigned for the design goals of a control system.

III. Determination of Equivalent Dominant Poles and Zeros from a High-Order Model

In the design of high performance control systems, quite often several specifications are assigned as design goals, and the corresponding dominant poles and zeros are required. This is a problem of a high-order transfer function fitting using industrial specifications. Shieh *et al.* [8], [9] have developed an original synthesis technique to fit a second-order transfer function based on three industrial specifications. The Newton-Raphson multidimensional method [10] was applied to solve the resulting nonlinear simultaneous equations that can be converted to a single variable quadratic equation. However, it is well known that the Newton-Raphson method will only con-

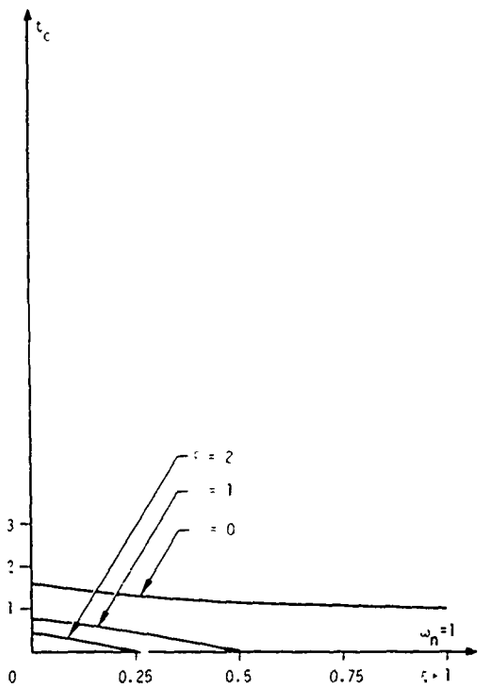


Fig. 10. Relationships among t_c, ξ, ω_n, and τ shown in (10a).

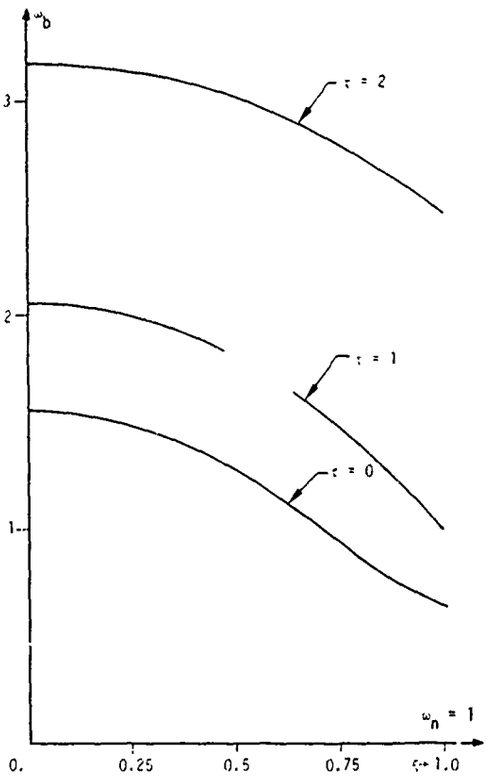


Fig. 11 Relationships among ω_b, ξ, ω_n, and τ shown in (13)

verge for a small range of starting values or the initial guesses. It is also known that high-order nonlinear equations have many solutions that depend heavily on the initial guess used. For general nonlinear equations that cannot be converted to a single variable equation, the Newton-Raphson numerical method may not converge to the desired solution using arbitrary initial guesses. In this paper, the original synthesis method [8], [9] is extended for modeling a high-order transfer function using many industrial specifications; and an analytical method is proposed for the estimation of the good starting values. T' us the desired dominant poles and zeros can be determined from the identified transfer function. The method can be well illustrated using the following example.

Suppose that the poles and zeros that represent the following given industrial specifications are required to be determined.

Type "1" system		(14a)
ω_c	the gain crossover frequency = 4.7 rad/s	(14b)
ϕ_m	the phase margin = 45.6°	(14c)
M_p	the maximum value of the closed-loop frequency response = 1.5	(14d)
ω_p	the peak value frequency = 3.5 rad/s	(14e)
ω_b	the bandwidth of the closed-loop frequency response = 6.5 rad/s.	(14f)

The assignments of the specifications in (14) closely follow the rules shown in (1). Therefore, the conflicted assignments can be avoided. The first two are the open-loop specifications, while the others are the closed-loop ones. Three equivalent poles and two equivalent zeros that represent the assigned specifications in (14) can be determined. The third-order model is

$$\frac{Y(s)}{R(s)} = T(s) = \frac{K(s+z_1)(s+z_2)}{(s^2+2\xi\omega_n s+\omega_n^2)(s+p)} = \frac{b_1 s^2 + b_2 s + b_3}{s^3 + a_1 s^2 + a_2 s + a_3} \tag{15a}$$

where K, p, ξ, ω_n, z_1, and z_2 or the corresponding a_i and b_i are unknown constants to be determined. Because the system is a type "1" system, the final value of the unit-step response of the system in (15a) is unity or

$$Y(t)\big|_{t\to\infty} = \lim_{s\to 0} s \cdot R(s)Y(s) = \lim_{s\to 0} s \cdot \left(\frac{1}{s}\right)\left(\frac{b_1 s^2 + b_2 s + b_3}{s^3 + a_1 s^2 + a_2 s + a_3}\right) = \frac{b_3}{a_3} = 1 \tag{15b}$$

or $$a_3 = b_3. \tag{15c}$$

As a result, (15a) can be simplified as

$$\frac{Y(s)}{R(s)} = T(s) = \frac{b_1 s^2 + b_2 s + a_3}{s^3 + a_1 s^2 + a_2 s + a_3}. \tag{15d}$$

The open-loop transfer function $G(s)$ is

$$G(s) = \frac{b_1 s^2 + b_2 s + a_3}{s[s^2 + (a_1 - b_1)s + (a_2 - b_2)]}. \tag{15e}$$

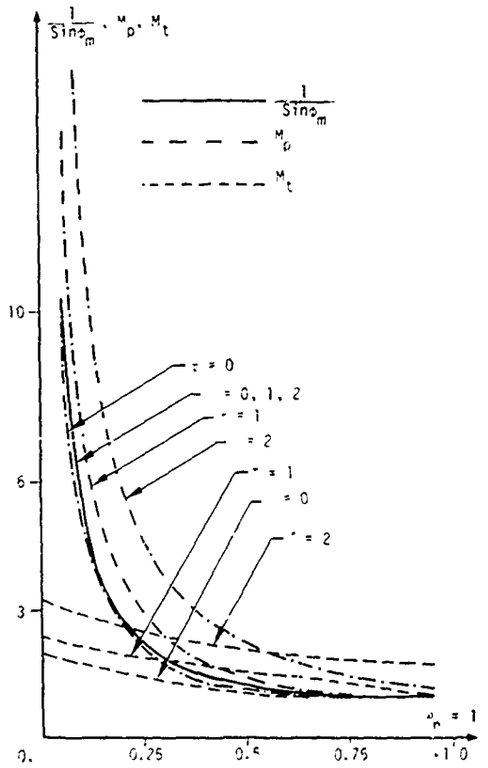


Fig. 12. Relationships among M_p, M_t, and $1/\sin\phi_m$ shown in (1).

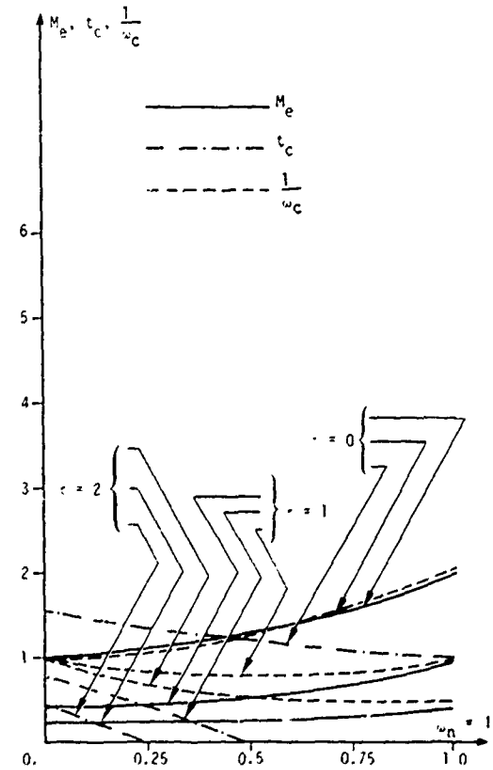


Fig. 13. Relationships among M_e, t_c, and $1/\omega_c$ shown in (1).

Following the definitions shown in (14), we can construct a set of nonlinear equations $f_i(a_1, a_2, a_3, b_1, b_2) = 0$ for $i = 1, 2, \cdots, 5$.

The definition of ω_c is

$$|G(j\omega_c)| = 1. \tag{16a}$$

The corresponding nonlinear equation is

$$f_1(a_1, a_2, a_3, b_1, b_2) = (a_1 - b_1)^2\omega_c^4 + [\omega_c^3 - (a_2 - b_2)\omega_c]^2 - (a_2 - b_1\omega_c^2)^2 - b_2^2\omega_c^2 = 0. \tag{16b}$$

The definition of ϕ_m can be expressed as

$$\phi_m = 180^\circ + \angle G(j\omega_c). \tag{17a}$$

The nonlinear equation is

$$f_2(a_1, a_2, a_3, b_1, b_2) = b_2\omega_c^2(a_1 - b_1) - (a_3 - b_1\omega_c^2)(\omega_c^2 - a_2 + b_2) - \tan\phi_m[(a_3 - b_1\omega_c^2)(a_1 - b_1)\omega_c + b_2\omega_c(\omega_c^2 - a_2 + b_2)] = 0. \tag{17b}$$

The definition of ω_b is known as

$$|T(j\omega_b)| = \frac{1}{\sqrt{2}}. \tag{18a}$$

The corresponding nonlinear equation is

$$f_3(a_1, a_2, a_3, b_1, b_2) = (a_3 - b_1\omega_b^2)^2 + b_2^2\omega_b^2 - \tfrac{1}{2}[(a_3 - a_1\omega_b^2)^2 + (\omega_b^3 - a_2\omega_b)^2] = 0. \tag{18b}$$

The definition of ω_p gives

$$\left.\frac{d|T(j\omega)|}{d\omega}\right|_{\omega=\omega_p} = 0. \tag{19a}$$

Following Higgins and Siegel's complex variable differential technique [7], we have the following nonlinear equation:

$$f_4(a_1, a_2, a_3, b_1, b_2) = [2a_1a_3\omega_p - 2a_1^2\omega_p^3 - (a_3 - 3\omega_p^2)(-\omega_p^3 + a_2\omega_p)][(a_3 - b_1\omega_p^2)^2 + (b_2\omega_p)^2] + [-2a_3b_1\omega_p + 2b_1^2\omega_p^3 + b_2^2\omega_p][(a_3 - a_1\omega_p^2)^2 + (-\omega_p^3 + a_2\omega_p)^2] = 0. \tag{19b}$$

The definition of M_p is

$$|T(j\omega)|_{\omega=\omega_p} = M_p. \tag{20a}$$

The nonlinear equation is

$$f_5(a_1, a_2, a_3, b_1, b_2) = (a_3 - b_1\omega_p^2)^2 + b_2^2\omega_p^2 - M_p^2[(a_3 - a_1\omega_p^2)^2 + (\omega_p^3 - a_2\omega_p)^2] = 0. \tag{20b}$$

Equations (16)-(20) are a set of high-order nonlinear simultaneous equations which are very difficult to solve. The Newton-Raphson method, which is available in most digital computers

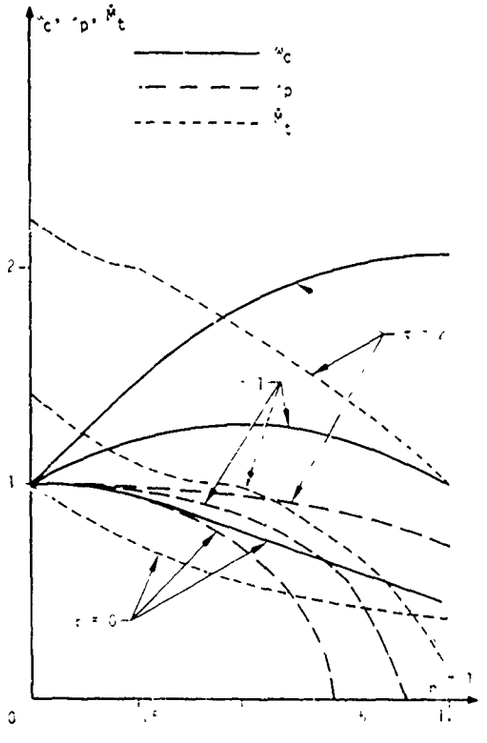


Fig. 14. Relationships among ω_c, ω_p, and M_t shown in (1).

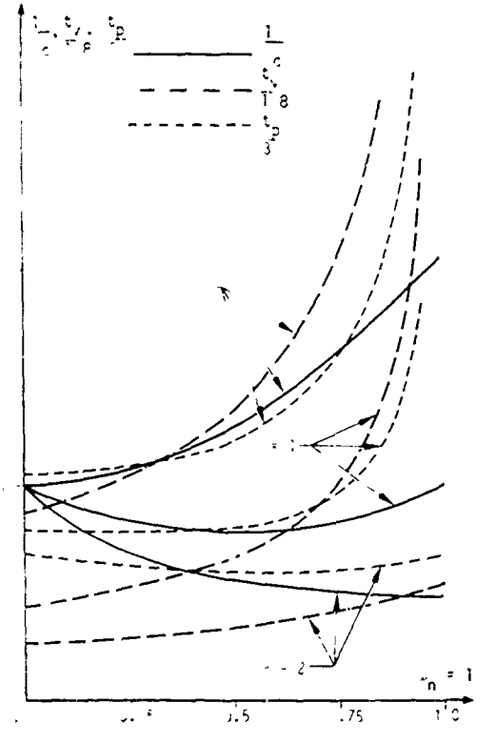


Fig. 15. Relationships among t_r/ω_c, $t_r/1.8$, and $t_p/3$ shown in (1).

[11], is used to solve the nonlinear equations. To obtain the desired solution, and to improve the speed of convergence of the numerical method, we have to establish a set of good starting values. From the developed analytical expressions of various specifications or the working curves in this paper, we can determine the corresponding two poles and one zero using $M_p = 1.5$ and $\omega_p = 3.5$. From the rule of the thumb in (1) we observe that the M_p and ω_p have indirectly included the approximated respective ϕ_m and ω_c. The procedures are shown in the following steps.

Step 1: Determine the normalized dominant poles or the ξ in (4a) using the curve drawn in Fig. 2, having $\tau = 0$. From the curve ($\tau = 0$) we read the damping ratio $\xi = 0.35$. The normalized dominant poles and the dominant poles with $\omega_p = \omega_n = 3.5$ are

$$s_1^* = 0.35 + j0.9368 \qquad s_1 = 1.225 + j3.2786$$
$$s_2^* = -0.35 - j0.9368 \qquad s_2 = 1.225 - j3.2786. \tag{21a}$$

The second-order model is

$$T_2^*(s) = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} = \frac{12.25}{s^2 + 2.45s + 12.25}. \tag{21b}$$

Step 2: Determine a dominant zero using the specification $\omega_b = 6.5$ in (14f). The modified second-order model becomes

$$T_2(s)^{**} = \frac{b_1 s + \omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} = \frac{b_1 s + 12.25}{s^2 + 2.45s + 12.25}. \tag{21c}$$

The b_1 can be easily determined by using the definition of ω_b in (18a):

$$b_1 = 3.1781. \tag{21d}$$

Thus a low-order dominant model is determined. However a third-order model is required. An extra pole and a nearby zero are inserted into the second-order model in (21c) to obtain an approximate third-order model, or

$$T_3^*(s) = \frac{(b_1^* s + \omega_n^2)(1.1s + 10\xi\omega_n)}{(s^2 + 2\xi\omega_n s + \omega_n^2)(s + 10\xi\omega_n)}$$
$$= \frac{3.49591s^2 + 52.406725s + 150.0625}{s^3 + 14.7s^2 + 42.2625s + 150.0625}. \tag{21e}$$

Using the coefficients in (21e) as initial guesses, $a_1^* = 14.7$, $a_2^* = 42.2625$, $a_3^* = 150.0625$, $b_1^* = 3.49591$, and $b_2^* = 52.406725$, and applying the Newton-Raphson method [11] to solve the nonlinear equations in (16) through (20) yields the desired solutions: $a_1 = 4.267162$, $a_2 = 20.58799$, $a_3 = 29.806197$, $b_1 = 3.188355$, and $b_2 = 15.561058$, at 10th iteration with the error tolerance of 10^{-6}. The desired transfer function is

$$T_3(s) = \frac{3.188355s^2 + 15.561058s + 29.806197}{s^3 + 4.267162s^2 + 20.58799s + 29.806197}. \tag{22}$$

The dominant poles and zeros, which represent the assigned industrial specifications, are determined from the poles P_i and zeros z_i in (22):

$$P_1 = 1.849412756$$
$$P_2 = 1.208824622 + j3.828226318$$
$$P_3 = 1.208824622 - j3.828226318 \tag{23a}$$

and

$$z_1 = 4.880591402 + j3.68424378$$
$$z_2 = 4.880591402 - j3.68424378. \tag{23b}$$

When the distribution of the poles and zeros of a high-order transfer function is known and the reduced-order transfer function that consists of equivalent dominant poles and zeros is required, it is a model reduction problem. Recently, various model reduction methods [12]-[15] have been proposed in the frequency domain. However, their reduced models [12]-[15], do not keep the assigned industrial specifications, which are obtained from the original system. The preservation of the exact frequency-domain specifications is essential in the design of filters and compensators using frequency-domain methods [5], [6], such as the Nyquist, Bode, and Nichols chart methods. This proposed method can overcome the shortcomings of the existing model reduction methods. The frequency-response data at ω_p, ω_b, ω_c, and ω_π (the phase crossover frequency of the open-loop system for the use of the gain margin [5], [6] are considered as the dominant frequency-response data. If some of these data are assigned to determine the corresponding reduced-order model, the equivalent dominant poles and zeros can be determined from the reduced-order model that consists of the exact industrial specifications assigned.

IV. Conclusion

A second-order transfer function with two poles and one finite zero has been used to derive the analytical and graphical expressions of various industrial specifications. For a few assigned industrial specifications, the corresponding two dominant poles and one dominant zero can be determined from the identified transfer function. The generalized second-order model has been used to verify the rule of the thumb proposed by Axelby. It has been observed that the accuracy of the rule of the thumb depends on the range of the damping ratio and the zero location. From the developed graphical expressions, a set of meaningful industrial specifications can be chosen and assigned as the design goals for the filter and compensator designs. A dominant frequency-response data matching method has been developed to construct a low-order transfer function using the assigned industrial specifications that are obtained from a given high-order system. Thus the equivalent dominant poles and zeros of a high-order system can be determined from the identified low-order transfer function that has the exact industrial specifications assigned.

More over, the proposed method in this paper has been successfully applied to redesign the compensators of a stabilized pitch control system of a real semiactive terminal homing missile [16]. The overall system characteristics of the redesigned missile [17] match those of the lower ordered model obtained from assigned industrial specifications.

References

[1] A. L. Greensite, *Elements of Modern Control Theory*. New York: Spartan Books, pp. 33-53, 1970.

[2] J. E. Gibson and Z. V. Rekasius, "A set of standard specifications for linear automatic control systems, "*AIEE Trans. Application and Industry*, pp. 65-77, May 1961.

[3] G. S. Axelby, "Practical methods of determining feedback control loop performance," in *Proc. 1st IFAC*, pp. 68-74, 1960.

[4] J. G. Truxal, *Control System Synthesis*. New York: McGraw-Hill, 1955, pp. 76-87.

[5] K. Ogata, *Modern Control Engineering*. Englewood Cliffs, NJ: Prentice-Hall, 1970, pp 216-258.

[6] B. C. Kuo, *Automatic Control Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1975, pp. 459-500.

[7] T. J. Higgins and C. M. Siegel, "Determination of the maximum modulus, or of the specified gain, of a servomechanism by complex-differentiation," *IEEE Trans. Application and Industry*, pp. 467-468, 1953.

[8] L. S. Shieh, "An algebraic approach to system identification and compensator design," Ph.D. dissertation, Univ. of Houston, Houston, TX, 1970.

[9] C. F. Chen and L. S. Shieh, "An algebraic method for control system design," *Int. J. Contr.*, vol. 11. pp. 717-739, 1970.

[10] B. Carnahan, H. A. Luther, and J. O. Wilkes, *Applied Numerical Methods*. New York: Wiley, 1969, pp. 319-329.

[11] IBM S/370-360 Reference Manual IMSL, The International Mathematical and Statistical Library.

[12] C. F. Chen and L. S. Shieh, "A novel approach to linear model simplification," *Int. J. Contr.*, vol. 8, no. 6, pp. 561-570, 1968.

[13] L. S. Shieh and M. J. Goldman, "A mixed Cauer form for linear system reduction," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, pp. 584-588, Nov. 1974.

[14] M. F. Hutton and B. Friedland, "Routh approximations for reducing order of linear, time-invariant systems," *IEEE Trans. Automatic Contr.*, vol. AC-20, pp. 329-337, June 1975.

[15] Y. Shamash, "Linear system reduction using pade approximation to allow retention of dominant models," *Int. J. Contr.*, vol. 21, pp. 257-272, 1975.

[16] J. T. Bosley, "Digital realization of the T-6 missile analog autopilot," Final Rep., U.S. Army Missile Command, DAAK40-77-C-0048, TGT-001, May 1977.

[17] M. Datta-Barua, "Redesigning the stabilized pitch control system of a semiactive terminal homing missile." M. S. thesis. Univ. of Houston, Houston, TX, 1978.

PROCEEDINGS

OF THE INSTITUTION OF ELECTRICAL ENGINEERS

Control & Science

Synthesis of optimal block controllers for multivariable control systems and its inverse optimal-control problem

Y.J. Wei, M.Sc., and Prof. L.S. Shieh, M.Sc., Ph.D.

Indexing terms: Multivariable control systems, Control-system synthesis, Optimal control

Abstract

A new method is presented to synthesise optimal block controllers for a class of multivariable control systems represented by the block companion form. The reverse process of obtaining the optimal block controller is used to determine the block-weighting matrices of the quadratic performance index from prescribed control specifications.

1 Introduction

The accurate description of linear time-invariant systems in the time domain may result in m nth-degree coupled differential equations, or an nth-degree matrix differential equation with $m \times m$ matrix coefficients[1] as

$$\sum_{i=1}^{n+1} A_i D^{i-1} x = u \tag{1a}$$

$$y = \sum_{i=1}^{n} C_i D^{i-1} x \tag{1b}$$

and

$$D^{i-1}x(0) = \alpha_i, \quad i = 1, 2, \ldots, n \tag{1c}$$

where y is an $m \times 1$ output vector, u is an $m \times 1$ input vector and x is an $m \times 1$ state vector. A_i and C_i are $m \times m$ matrix coefficients, and the differential operator $D = d/dt$. When each initial vector α_i is an $m \times 1$ null vector, the corresponding frequency-domain representation of eqn. 1 is an nth-degree matrix transfer function written as

$$Y(s) = T(s)U(s) \tag{2a}$$

where $Y(s)$ and $U(s)$ are the $m \times 1$ output vector and the $m \times 1$ input vector, respectively, and the matrix transfer function $T(s)$ is

$$T(s) = N_r(s)D_r^{-1}(s) = D_l^{-1}(s)N_l(s) \tag{2b}$$

The matrix polynomials $D_r(s)$ and $N_r(s)$ with appropriate size are right coprime, $D_l(s)$ and $N_l(s)$ left coprime. Let us define

$$\begin{aligned} D_r(s) &= I_m s^n + A_n s^{n-1} + \ldots + A_2 s + A_1 \\ N_r(s) &= C_n s^{n-1} + C_{n-1} s^{n-2} + \ldots + C_2 s + C_1 \end{aligned} \tag{3}$$

where A_i and C_i are $m \times m$ constant matrices. The corresponding first-degree state equation in the controllable phase-variable block form or in the controllable block companion form is

$$\dot{X} = AX + Bu \tag{4a}$$

$$y = CX; x(0) = X_0 \tag{4b}$$

where

$$A = \begin{bmatrix} O_m & I_m & O_m & . & O_m \\ O_m & O_m & I_m & . & O_m \\ . & . & . & . & . \\ -A_1 & -A_2 & -A_3 & . & -A_n \end{bmatrix}, \quad B = \begin{bmatrix} O_m \\ O_m \\ . \\ I_m \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ . \\ X_n \end{bmatrix} \tag{4c}$$

$$C = [C_1 \quad C_2 \quad \ldots \quad C_n] \tag{4d}$$

The block elements A_i, O_m, I_m and C_i are $m \times m$ constant matrices, $m \times m$ null matrix, $m \times m$ identity matrix and $m \times m$ constant matrices, respectively. The vector X consists of n blocks (X_i, $i = 1, 2, \ldots, n$) and each $m \times 1$ block X_i consists of m state variables. In this paper, we define the vector X as a block vector. Because the state equation in eqn. 4 is formulated in the phase-variable block form, the X is defined as a vector in the phase-variable block co-ordinate. As a result, the $X(0)$ is an initial block vector. From a conventional viewpoint, the same vector X is viewed as a vector with nm state variables in a general co-ordinate. Therefore, the same state equation in eqn. 4 is viewed as a state equation in a general co-ordinate. In this paper, all the derivations are based on the state equation in the phase-variable block co-ordinate rather than a general co-ordinate.

The objectives of this paper are described as follows:

(*a*) Obtain the optimal block-cont law $u = -R^{-1}B^TPX = -KX$ (where the feedback-gain matrix $K = R^{-1}B^TP$ consists of $m \times m$ block elements K_i, $i = 1, \ldots, n$) to minimise the quadratic performance index

$$J = \frac{1}{2}\int_0^\infty [X^TQX + u^TRu]dt \tag{5a}$$

for the dynamic system formulated in the phase-variable block co-ordinate in eqn. 4. The T designates transpose, the weighting matrix R is an assigned $m \times m$ positive-definite matrix, and the block-weighting matrix Q is an assigned $nm \times nm$ nonnegative definite-symmetric matrix with $m \times m$ block elements $Q_{i,j} = Q_{j,i}^T$, or

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & \ldots & Q_{1n} \\ Q_{21} & Q_{22} & \ldots & Q_{2n} \\ . & . & \ldots & . \\ Q_{n1} & Q_{n2} & \ldots & Q_{nn} \end{bmatrix} = Q^T \tag{5b}$$

The $nm \times nm$ matrix P is the positive-definite solution of the steady-state Riccati equation[2]

$$PA + A^TP + Q - PBR^{-1}B^TP = O_{nm} \tag{5c}$$

The same P can be also solved from the following canonical form:[2]

$$\begin{bmatrix} \dot{X} \\ \dot{G} \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} X \\ G \end{bmatrix},$$

$$G(\infty) = PX(\infty) = O_{nm \times 1}, X(0) = X_0 \tag{5d}$$

Paper 8267C, received 3rd November 1978 and in revised form 24th January 1979

Mr. Wei and Prof. Shieh are with the Department of Electrical Engineering, University of Houston, Central Campus, Houston, Texas 77004, USA

0020–3270/79/050449 + 08 $01·50/0

It is noted that, if the pair $[A, B]$ is controllable and the pair $[A, L]$ is observable (where $Q = LL^T$), then the closed-loop system is not only optimal but stable.

(*b*) Determine the block-weighting matrices Q and R of the quadratic performance index in eqn. 5*a* if the optimal block controller K is assigned or if the closed-loop poles (or the equivalent control specifications[3]) of the optimal controlled system are prescribed.

2 Linear optimal-block-regulator problem

In the conventional synthesis of the linear-regulator problem, the state equation in eqn. 4 is viewed as a state equation in a general co-ordinate. An optimal control law is then derived by solving eqns. 5*c* or 5*d*. In this paper, the state equation in eqn. 4 is considered as a state equation in the phase-variable block co-ordinate. The optimal-block-control law is derived as follows.

Expanding eqn. 4 and adding a trivial identity yields

$$\begin{aligned} X_1 &= X_1 \\ \dot{X}_1 &= X_2 \\ \ddot{X}_1 &= X_3 = \dot{X}_2 \\ &\cdots\cdots \\ X^{(n)} &= X_n = -A_1X_1 - A_2X_2 - \ldots - A_nX_n + u \end{aligned} \tag{6a}$$

Rewriting the last equation in eqn. 6*a* gives

$$u = A_1X_1 + A_2\dot{X}_1 + \ldots + A_nX_1^{(n-1)} + X_1^{(n)} \tag{6b}$$

Substituting eqn. 6 into eqn. 5*a*, we have an alternate form of the cost function as

$$F(X, u) = F(X_1, \dot{X}_1, \ldots, X_1^{(n)}) = F(X^*) = \tfrac{1}{2}X^{*T}Q^*X^* \tag{7}$$

where

$$Q^* = \begin{bmatrix} Q_{11}^* & Q_{12}^* & \ldots & Q_{1n}^* & A_1^TR \\ Q_{21}^* & Q_{22}^* & \ldots & Q_{2n}^* & A_2^TR \\ & & \ldots & & \\ Q_{n1}^* & Q_{n2}^* & \ldots & Q_{nn}^* & A_n^TR \\ RA_1 & RA_2 & \ldots & RA_n & R \end{bmatrix} = Q^{*T}, \quad X^* = \begin{bmatrix} X_1 \\ \dot{X}_1 \\ \\ X_1^{(n-1)} \\ X_1^{(n)} \end{bmatrix}$$

$$Q_{i,j}^* = Q_{i,j} + A_i^TRA_j = Q_{j,i}^{*T}$$

The $(n+1)m \times (n+1)m$ constant matrix Q^* is a block weighting matrix with $m \times m$ block elements. Applying the gradient matrix operations[4] to the quadratic cost function in eqn. 7 yields

$$\begin{aligned} F_{X_1} &= [I_m \quad O_m \quad \ldots \quad O_m]Q^*X^* \\ \frac{d}{dt}F_{\dot{X}_1} &= [O_m \quad I_m \quad \ldots \quad O_m]Q^*\dot{X}^* \\ &\cdots\cdots \\ \frac{d^n}{dt^n}F_{X_1^{(n)}} &= [O_m \quad O_m \quad \ldots \quad I_m]Q^*X^{*(n)} \end{aligned} \tag{8}$$

Substituting eqn. 8 into the following Euler's equation[5]

$$F_{X_1} - \frac{d}{dt}F_{\dot{X}_1} + \frac{d^2}{dt^2}F_{\ddot{X}_1} - \ldots + (-1)^n\frac{d^n}{dt^n}F_{X_1^{(n)}} = O_{m\times 1} \tag{9}$$

we have

$$D_1X_1 + D_2\dot{X}_1 + D_3X_1^{(2)} + \ldots + D_{2n+1}X_1^{(2n)} = O_{m\times 1} \tag{10a}$$

where

$$[D_1 \quad D_2 \quad D_3 \quad \ldots \quad D_{2n+1}] = [I_m \quad -I_m \quad I_m \quad \ldots \quad (-1)^nI_m] \times$$

$$\begin{bmatrix} Q_{11}^* & Q_{12}^* & Q_{13}^* & \ldots & Q_{1n}^* & A_1^TR & O_m & \ldots & O_m & O_m & O_m \\ O_m & Q_{21}^* & Q_{22}^* & \ldots & Q_{2,n-1}^* & Q_{2,n}^* & A_2^TR & \ldots & O_m & O_m & O_m \\ O_m & O_m & Q_{31}^* & \ldots & Q_{3,n-2}^* & Q_{3,n-1}^* & Q_{3,n}^* & \ldots & O_m & O_m & O_m \\ \cdot & \cdot & \cdot & \ldots & \cdot & \cdot & \cdot & \ldots & \cdot & \cdot & \\ O_m & O_m & O_m & \ldots & Q_{n,1}^* & Q_{n,2}^* & Q_{n,3}^* & \ldots & Q_{n,n}^* & A_n^TR & O_m \\ O_m & O_m & O_m & \ldots & O_m & RA_1 & RA_2 & \ldots & RA_{n-1} & RA_n & R \end{bmatrix} \tag{10b}$$

Expanding eqn. 10*b* we have

$$\begin{aligned} D_1 &= Q_{11}^* = Q_{11} + A_1^TRA_1 \\ D_2 &= Q_{12}^* - Q_{21}^* = Q_{12} + A_1^TRA_2 - Q_{21} - A_2^TRA_1 \\ &\cdots\cdots \\ D_{2n+1} &= R \end{aligned} \tag{10c}$$

Taking the Laplace transform of eqn. 10*a* and neglecting the initial conditions we have the matrix polynomial $D(s)$:

$$\begin{aligned} D(s)X_1(s) &= [D_{2n+1}s^{2n} + D_{2n}s^{2n-1} + \ldots \\ &\quad + D_2s + D_1]X_1(s) = O_{m\times 1} \end{aligned} \tag{11}$$

where $D_{2k+1} = D_{2k+1}^T$, $k = 0, 1, \ldots, n$ and $D_{2k} = -D_{2k}^T$, $k = 1, 2, \ldots, n$. It is well known that the poles of the state equation in eqn. 5*d* are symmetrically distributed about the origin in the s-plane, so are the roots of the determinant of the matrix polynomial $D(s)$ in eqn. 11. Performing the spectral factorisation[6,7] of the matrix polynomial $D(s)$ results in a stable matrix polynomial $\Delta(s)$ and an unstable matrix polynomial $\Delta(-s)$, i.e.

$$D(s) = F^T\Delta(-s)^T\Delta(s)F \tag{12}$$

where

$$R = F^TF = D_{2n+1}$$

and

$$\Delta(s) = I_ms^n + E_ns^{n-1} + \ldots + E_2s + E_1$$

The required optimal-block-control law is then obtained from eqns. 6*b* and 12 as

$$u = [K_1 \quad K_2 \ldots K_n]X \tag{13}$$

where

$$K_i = A_i - E_i, \quad i = 1, 2, \ldots, n$$

When the given system is not in a phase-variable block form, a newly developed algorithm shown in Appendix 8 can be applied to obtain a block linear transformation that transforms a class of state equations in a general co-ordinate into the phase-variable block co-ordinate. Thus the proposed method can be applied to determine the optimal block controller.

3 Inverse optimal control problem

Given a set of prescribed closed-loop poles, or equivalent control specifications,[3] we wish to determine the weighting matrices Q and R of the quadratic performance index in eqn. 5*a* by which the controlled feedback system has prescribed closed-loop poles and the feedback-control law is optimal. This is an inverse optimal-control problem. Kalman[8] initiated the inverse problem for a linear time-invariant single-input system. Chang,[9] Tyler and Tuteur[10] have studied the problem via the root-locus method, while Molinari,[11] and Anderson and Shannon[12] have investigated the problem for a multivariable system. All the developed methods are based on the system equation formulated in a general co-ordinate rather than in a phase-variable block co-ordinate. Since the multivariable dynamic system is formulated in a matrix differential equation, it is more natural to investigate the problem in the phase-variable block co-ordinate than that in the general co-ordinate.

It is well known that a feedback-gain matrix can always be obtained to give a system with prescribed closed-loop poles if a system is controllable. However, the feedback controller may not be optimal. In this paper we determine the block-weighting matrices Q and R of the quadratic performance index by which the feedback controller not only provides the controlled system with prescribed closed-loop poles but also performs optimally. The steps involved are described as follows:

Step 1

Define a characteristic matrix polynomial $\Delta(s)$ of the desired closed-loop system whose matrix coefficients consist of some unknown parameters (for example, the damping ratio ξ and the undamped natural angular frequency ω_n etc.) to be adjusted. The $\Delta(s)$ is

$$\Delta(s) = I_ms^n + E_ns^{n-1} + \ldots + E_2s + E_1 \tag{14a}$$

If the desired characteristic polynomial of the closed-loop system is

$$[d(s)]^m = (s^n + d_ns^{n-1} + \ldots + d_2s + d_1)^m \tag{14b}$$

where $d(s)$ is a polynomial whose coefficients consist of adjustable parameters. The characteristic matrix polynomial becomes

$$\Delta(s) = d(s)I_m = I_m s^n + d_n I_m s^{n-1} + \ldots + d_2 I_m s + d_1 I_m \tag{14c}$$

where

$$E_i = d_i I_m$$

Step 2

Construct a matrix polynomial $D(s)$ using $\Delta(s)$ in eqn. 14

$$\begin{aligned} D(s) &= D_{2n+1}s^{2n} + D_{2n}s^{2n-1} + \ldots + D_2 s + D_1 \\ &= F^T\Delta^T(-s)\Delta(s)F \\ &= F^T[I_m s^{2n} + (E_n - E_n^T)s^{2n-1} \\ &\quad + (E_{n-1} - E_n^T E_n + E_{n-1}^T)s^{2n-2} + \ldots + E_1^T E_1]F \end{aligned} \tag{15}$$

where $D_{2n+1} = F^T F = R$ is a weighting matrix to be determined.

Step 3

Solve the block weighting matrices Q and R from eqns. 10 and 15 in terms of adjustable parameters, or

$$\begin{aligned} D_{2n+1} &= F^T F \\ D_{2n} &= RA_n - A_n^T R = F^T(E_n - E_n^T)F \\ &\cdots\cdots \\ D_2 &= Q_{12} + A_1^T RA_2 - Q_{21} - A_2^T RA_1 \\ &= F^T(E_1^T E_2 - E_2^T E_1)F \\ D_1 &= Q_{11} + A_1^T RA_1 = F^T E_1^T E_1 F \end{aligned} \tag{16}$$

Step 4

Determine the required block weighting matrices Q and R by adjusting the assigned unknown parameters such that R is positive definite and Q is nonnegative definite symmetric.

The procedures can be well illustrated by the following gas-turbine example.

4 An illustrative example

Consider the following linearised two-shaft gas-turbine model.[13-15]

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} -1{\cdot}268 & -0{\cdot}04528 & 1{\cdot}498 & 951{\cdot}5 \\ 1{\cdot}002 & -1{\cdot}957 & 8{\cdot}52 & 1240 \\ 0 & 0 & -10 & 0 \\ 0 & 0 & 0 & -100 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 10 & 0 \\ 0 & 100 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{17a}$$

and

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \tag{17b}$$

The state equation in eqn. 17 is a system formulated in a general co-ordinate. To apply the proposed method, a block-linear-transformation matrix T is determined from the newly developed method shown in Appendix 8. The block linear transformation is

$$z = TX \tag{18}$$

where

$$T = \left[\begin{array}{cc:cc} 14{\cdot}98 & 95150 & 0 & 0 \\ 85{\cdot}2 & 124000 & 0 & 0 \\ \hdashline 18{\cdot}5671 & -2622{\cdot}1 & 10 & 0 \\ -0{\cdot}005214 & 136{\cdot}829 & 0 & 100 \end{array}\right]$$

and X is in the phase-variable block co-ordinate and consists of two block vectors (X_i, $i = 1, 2$) and each vector X_i consists of two state variables ($x_{i,j}$, $j = 1, 2$, $i = 1, 2$). The state equation in the phase-variable block co-ordinate is

$$\begin{bmatrix} \dot{x}_{1,1} \\ \dot{x}_{1,2} \\ \dot{x}_{2,1} \\ \dot{x}_{2,2} \end{bmatrix} = \left[\begin{array}{cc:cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hdashline -18{\cdot}5671 & 2622{\cdot}1 & -11{\cdot}8567 & 262{\cdot}21 \\ 0{\cdot}005214 & -136{\cdot}829 & 0 & -101{\cdot}368 \end{array}\right] \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ x_{2,1} \\ x_{2,2} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \hdashline 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{19a}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \left[\begin{array}{cc:cc} 14{\cdot}98 & 95150 & 0 & 0 \\ 85{\cdot}2 & 124000 & 0 & 0 \end{array}\right] \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \hdashline x_{2,1} \\ x_{2,2} \end{bmatrix} \tag{19b}$$

where

$$A_1 = \begin{bmatrix} 18{\cdot}5671 & -2622{\cdot}1 \\ -0{\cdot}005214 & 136{\cdot}829 \end{bmatrix}, A_2 = \begin{bmatrix} 11{\cdot}8567 & -262{\cdot}21 \\ 0 & 101{\cdot}368 \end{bmatrix} \tag{19c}$$

$$C_1 = \begin{bmatrix} 14{\cdot}98 & 95150 \\ 85{\cdot}2 & 124000 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{19d}$$

It is required to determine two optimal block controllers for the gas-turbine system by using

(*a*) assigned weighting matrices Q and R of the quadratic performance index

(*b*) assigned control specifications.

The procedures are described as follows:

(*a*) *Optimal-block-controller design via assigned weighting matrices*

The cost function of the state equation in the original co-ordinate in eqn. 17 is

$$J = \frac{1}{2}\int_0^\infty \{z^T\bar{Q}z + u^T Ru\}\,dt \tag{20}$$

where $Q = I_4$ and $R = I_2$ that were suggested by Tiwari *et al.*[15] The corresponding cost function in the phase-variable block co-ordinate is

$$J = \frac{1}{2}\int_0^\infty [X^T QX + u^T Ru]\,dt \tag{21}$$

where $R = I_2$ and

$$Q = T^T\bar{Q}T = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \left[\begin{array}{cc:cc} 7828{\cdot}1776 & 11941461{\cdot}5 & 185{\cdot}671 & -{\cdot}521389 \\ 11941461{\cdot}5 & 24436416630 & -26221 & 13682{\cdot}9 \\ \hdashline 185{\cdot}671 & -26221 & 100 & 0 \\ -{\cdot}521389 & 13682{\cdot}9 & 0 & 10000 \end{array}\right]$$

From eqn. 10 we have

$$[D_1\ D_2 \ldots D_5] = [I_2\ \ -I_2\ \ I_2] \begin{bmatrix} Q_{11}^* & Q_{12}^* & A_1^T R & O_2 & O_2 \\ O_2 & Q_{21}^* & Q_{22}^* & A_2^T R & O_2 \\ O_2 & O_2 & RA_1 & RA_2 & R \end{bmatrix} \tag{22}$$

By expanding eqn. 22, $D(s)$ in eqn. 11 becomes

$$\begin{aligned} D(s) &= D_5 s^4 + D_4 s^3 + \ldots + D_1 \\ &= Rs^4 + (RA_2 - A_2^T R)s^3 \\ &\quad + (RA_1 + A_1^T R - Q_{22} - A_2^T RA_2)s^2 \\ &\quad + (Q_{12} + A_1^T RA_2 - Q_{21} - A_2^T RA_1)s \\ &\quad + (Q_{11} + A_1^T RA_1) = O_2 \end{aligned} \tag{23}$$

where

$$D_5 = I_2, D_4 = \begin{bmatrix} 0 & -262{\cdot}21 \\ 262{\cdot}21 & 0 \end{bmatrix}$$

$$D_3 = \begin{bmatrix} -2{\cdot}03447 & 486{\cdot}84 \\ 486{\cdot}84 & -88755{\cdot}96 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 0 & 52440{\cdot}924 \\ -52440{\cdot}924 & 0 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} 8172{\cdot}92 & 11892776 \\ 11892776 & 2{\cdot}44433 \times 10^{10} \end{bmatrix}$$

Performing the spectral factorisation[7] on the $D(s)$ gives

$$\Delta(s) = I_2 s^2 + E_2 s + E_1 \tag{24}$$

where

$$E_2 = \begin{bmatrix} 17{\cdot}2396 & -261{\cdot}906 \\ 0{\cdot}30451 & 576{\cdot}845 \end{bmatrix} \text{ and } E_1 = \begin{bmatrix} 46{\cdot}925 & -3929{\cdot}922 \\ 77{\cdot}27215 & 156294{\cdot}2 \end{bmatrix}$$

From eqn. 13 we have the optimal block controllers in the block co-ordinate and original co-ordinate as

$$\begin{aligned} u &= [A_1 - E_1 \vdots A_2 - E_2]X \\ &= -\left[\begin{array}{cc:cc} 28{\cdot}3576 & -1307{\cdot}82 & 5{\cdot}3829 & 0{\cdot}3045 \\ 77{\cdot}2774 & 156157{\cdot}4 & 0{\cdot}304513 & 475{\cdot}476 \end{array}\right]X \end{aligned} \tag{25a}$$

$$\begin{aligned} &= [A_1 - E_1 \vdots A_2 - E_2]T^{-1}z \\ &= \left[\begin{array}{cc:cc} -0{\cdot}36296 & 0{\cdot}279346 & 0{\cdot}53829 & 0{\cdot}003045 \\ 0{\cdot}598572 & 0{\cdot}795425 & 0{\cdot}0304513 & 4{\cdot}75476 \end{array}\right]z \end{aligned} \tag{25b}$$

(*b*) *The optimal-block-controller design via assigned control specifications*

The design goals are specified as follows:

(i) static decoupling
(ii) final values of the unit-step responses are unity
(iii) peak time t_p that is the time required for the unit-step response to reach the first peak of the overshoot is near 0·01 s
(iv) maximum percentage overshoot is less than 10%.

To reach the first design goal, the characteristic matrix polynomial is defined as

$$\Delta(s) = I_2 s^2 + E_2 s + E_1 \tag{26}$$

where

$$E_2 = \begin{bmatrix} 2\xi\omega_n & 0 \\ 0 & 2\xi\omega_n \end{bmatrix} \text{ and } E_1 = \begin{bmatrix} \omega_n^2 & 0 \\ 0 & \omega_n^2 \end{bmatrix}$$

ξ (damping ratio) and ω_n (undamped natural frequency) are unknown parameters to be determined. To satisfy the third design goal we can estimate ω_n from the following rule of thumb in designs[16] as

$$\omega_n \simeq \frac{\pi}{t_p} \simeq \frac{3{\cdot}14}{0{\cdot}01} \simeq 300 \text{ rad/s} \tag{27a}$$

Also, from another rule of thumb,[16] we can estimate ξ to meet the fourth design goal as

$$\xi \cong -\frac{\ln M_p}{\pi} = -\frac{\ln 0{\cdot}1}{3{\cdot}14} \simeq 0{\cdot}75 \tag{27b}$$

The choices in eqn. 27 imply that the closed-loop poles have been assigned at

$$s_{1,2} = -\xi\omega_n \pm j\omega_n\sqrt{1-\xi^2} = -225 \pm j\,198{\cdot}43 \tag{27c}$$

From eqn. 26 $D(s)$ can be determined as

$$\begin{aligned} D(s) &= F^T\Delta^T(-s)\Delta(s)F \\ &= F^T F s^4 + (2\omega_n^2 - 4\xi^2\omega_n^2)F^T F s^2 + \omega_n^4 F^T F \\ &= Rs^4 + (RA_2 - A_2^T R)s^3 \\ &\quad + (RA_1 + A_1^T R - Q_{22} - A_2^T RA_2)s^2 \\ &\quad + (Q_{12} + A_1^T RA_2 - Q_{12}^T - A_2^T RA_1)s \\ &\quad + (Q_{11} + A_1^T RA_1) \end{aligned} \tag{28}$$

For simplicity, let $Q_{12} = Q_{21} = O_2$. Equating the matrix coefficients of the same power of eqn. 28, we obtain the following matrix equations:

$$(a)\ R = F^T F \tag{29a}$$

$$(b)\ RA_2 - A_2^T R = O_2 \tag{29b}$$

$$(c)\ RA_1 + A_1^T R - Q_{22} - A_2^T RA_2 = (2\omega_n^2 - 4\xi^2\omega_n^2)F^T F \tag{29c}$$

$$(d)\ A_1^T RA_2 - A_2^T RA_1 = O_2 \tag{29d}$$

$$(e)\ Q_{11} + A_1^T RA_1 = \omega_n^4 F^T F \tag{29e}$$

R is an $m \times m$ symmetric and positive-definite matrix which has $m(m+1)/2$ unknown elements to be determined. The left-hand-side matrices in eqns. 29*b* and 29*d* are skew-symmetric matrices. Expanding the matrix equations in eqns. 29*b* and 29*d* results in $m(m-1)$ simultaneous equations with $m(m+1)/2$ unknown variables in R. In general, there are an infinite number of solutions. However, if k independent simultaneous equations exist, and $k < m(m+1)/2$, then we can assume $[m(m+1)/2 - k]$ constants to solve k unknown variables in R. The choice of the assigned constants in R is a design freedom and a certain amount of experience is helpful. In this example, we assume R_{11}, which is the first leading diagonal element, is unity. Thus we can solve for R and F in eqn. 29*a* as

$$R = \begin{bmatrix} 1 & 2{\cdot}92934 \\ 2{\cdot}92934 & 51058{\cdot}01562 \end{bmatrix} = F^T F \tag{30}$$

where

$$F = \begin{bmatrix} 0{\cdot}999916 & 5{\cdot}737 \times 10^{-5} \\ 0{\cdot}0129466 & 225{\cdot}96 \end{bmatrix}$$

Note that R is a positive-definite matrix. From eqns. 30, 29*c* and 29*e* we can solve for Q_{11} and Q_{22} as

$$Q_{11} = \begin{bmatrix} \omega_n^4 - 345{\cdot}55808 & 2{\cdot}929341\omega_n^4 + 7{\cdot}629{\cdot}05 \\ 2{\cdot}929341\omega_n^4 + 7{\cdot}629{\cdot}05 & 51058{\cdot}0156\,\omega_n^4 - 7{\cdot}6070451 \times 10^8 \end{bmatrix} \tag{31a}$$

$$Q_{22} = \begin{bmatrix} 4\xi^2\omega_n^2 - 2\omega_n^2 - 140{\cdot}5813 & 2{\cdot}929341(4\xi^2\omega_n^2 - 2\omega_n^2) - 2844{\cdot}916 \\ 2{\cdot}929341(4\xi^2\omega_n^2 - 2\omega_n^2) - 2844{\cdot}916 & 51058{\cdot}0156(4\xi^2\omega_n^2 - 2\omega_n^2) - 524561317{\cdot}4 \end{bmatrix} \tag{31b}$$

Substituting $\omega_n = 300$ and $\xi = 0{\cdot}75$ into eqn. 31 yields positive-definite matrices Q_{11} and Q_{22}. Thus the optimal block controllers can be easily found in the block co-ordinate and in the original co-ordinate as

$$\begin{aligned} u &= [A_1 - E_1 \vdots A_2 - E_2]X \\ &= -\left[\begin{array}{cc:cc} 89981{\cdot}4329 & 2622{\cdot}1 & 438{\cdot}1433 & 262{\cdot}21 \\ 0{\cdot}005214 & 89863{\cdot}17 & 0 & 348{\cdot}632 \end{array}\right]X \end{aligned} \tag{32a}$$

$$= [A_1 - E_1 \vdots A_2 - E_2]\,T^{-1}z$$

$$= \left[\begin{array}{cc:cc} -1767{\cdot}7 & 1357{\cdot}37 & 43{\cdot}8143 & 2{\cdot}6221 \\ 1{\cdot}2182 & -0{\cdot}21391 & 0 & 3{\cdot}48632 \end{array}\right] z \qquad (32b)$$

The block-weighting matrix Q in the block co-ordinate and the weighting matrix $\bar{Q}$ in the original co-ordinate are

$$Q = \left[\begin{array}{cc:cc} 8{\cdot}1 \times 10^9 & 2{\cdot}37277 \times 10^{10} & 0 & 0 \\ 2{\cdot}37277 \times 10^{10} & 4{\cdot}13569 \times 10^{14} & 0 & 0 \\ \hdashline 0 & 0 & 11703{\cdot}42 & 31850{\cdot}2 \\ 0 & 0 & 31850{\cdot}2 & 80169820 \end{array}\right] \qquad (33a)$$

and

$$\bar{Q} = \left[\begin{array}{cc:cc} 3253160 & -2454610 & 47{\cdot}2383 & -2{\cdot}91217 \\ -2454610 & 1878440 & -33{\cdot}808 & -5{\cdot}9383 \\ \hdashline 47{\cdot}2383 & -33{\cdot}808 & 117{\cdot}0342 & 31{\cdot}8502 \\ -2{\cdot}91217 & -5{\cdot}9383 & 31{\cdot}8502 & 8016{\cdot}982 \end{array}\right] \qquad (33b)$$

It is noticed that[2] any arbitrarily prescribed closed-loop poles or control specifications may not result in a positive-definite matrix R and nonnegative-definite matrix Q. The constraints suggested by Anderson[2] should be satisfied. In addition, some realistic constraints to the amplitudes of the control signals, for example the limitations of the actuator amplitude and rate change of amplitude, should be also examined.

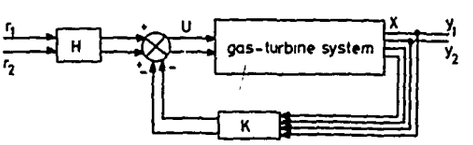


Fig. 1
Structure of designed system

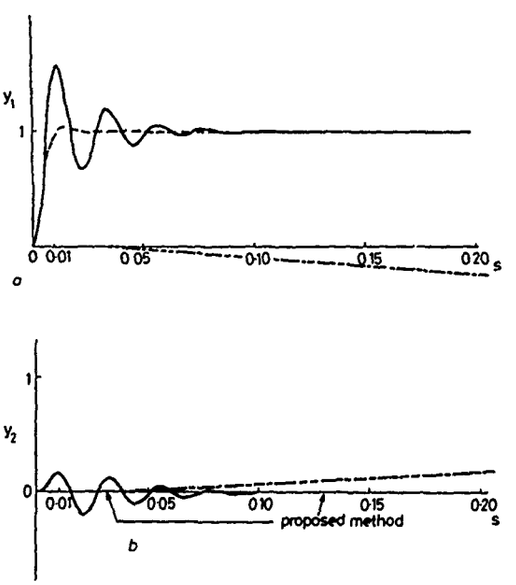


Fig. 2
Responses of various designed systems to a unit step in r_1

$r_1 = 1$
$r_2 = 0$
——— McMorran's method:
– – – proposed method: $\xi = 0{\cdot}75$; $\omega_n = 300$
– · – · proposed method: $\bar{Q} = I_4$; $R = I_2$
Tiwari's method: $\bar{Q} = I_4$; $R = I_2$

To achieve the first and second design goals we add a forward-gain matrix H as shown in Fig. 1. The H can be solved from the block C_1 in eqn. 19*d* or

$$H = \omega_n^2 \begin{bmatrix} 14{\cdot}98 & 95150 \\ 85{\cdot}2 & 124000 \end{bmatrix}^{-1} = \begin{bmatrix} -198{\cdot}42349 & 152{\cdot}258 \\ 0{\cdot}136336 & -0{\cdot}023971 \end{bmatrix} \qquad (34)$$

Thus the design system is

$$\begin{bmatrix} y_1(s) \\ y_2(s) \end{bmatrix} = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_1(s) \\ R_2(s) \end{bmatrix}$$

$$= \frac{90000}{s^2 + 450s + 90000} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_1(s) \\ R_2(s) \end{bmatrix} \qquad (35)$$

For this real nontrivial system the designed system is not only static decoupling but also complete noninteracting, and the final values of the unit-step responses are unity. The peak time is 0·014 s and the maximum percentage overshoot is 1%. The simulation curves for unit-step input are shown in Figs. 2 and 3. Comparing the design results of the proposed method with those of McMorran[14] and Tiwari *et al.*,[15] the present result gives less overshoot and less oscillatory responses.

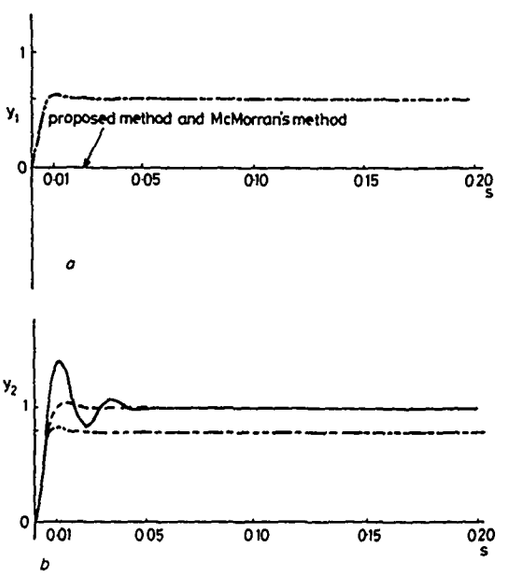


Fig. 3
Responses of various designed systems to a unit step in r_2

$r_1 = 0$
$r_2 = 1$
——— McMorran's method
– – – proposed method. $\xi = 0{\cdot}75$; $\omega_n = 300$
– · – · proposed method. $\bar{Q} = I_4$; $R = I_2$
Tiwari's method: $\bar{Q} = I_4$; $R = I_2$

5 Conclusion

A new method, based on a state equation in the phase-variable block co-ordinate, has been presented to determine the optimal block controllers for a class of multivariable systems. The reverse process of obtaining the optimal block controllers has been used to determine the weighting matrices of the quadratic performance index.

When a multivariable dynamic system is formulated in a matrix differential equation, the proposed method is more suitable for the determination of the optimal controllers than the conventional approach. Also, it is simpler to determine the weighting matrices than the conventional approaches. However, the proposed method is limited to a class of multivariable systems whose state equations can be formulated into matrix differential equations or the state equations in the block companion form.

6 Acknowledgments

The authors wish to express their gratitude for the valuable remarks and suggestions of the referees. This work was supported in part by US Army Missile Research and Development Command, DAAK 40-78-C-0017, DAAK 40-79-C-0061, and US Army Research Office DAAG 29-77-G-0143.

7 References

1 WELLS, D.A.: 'Lagrangian dynamics' (Schaum, New York, 1967), pp. 1–8
2 ANDERSON, B.D.O., and MOORE, J.B.: 'Linear optimal control', (Prentice-Hall, New Jersey, 1971), pp. 50–115
3 GIBSON, J.L., and REKASIUS, Z.V.: 'A set of standard specifications for linear automatic control systems', *Trans. Amer. Inst. Elect. Engrs.*, 1961, 80, pp. 65–77
4 KIRK, D.E.: 'Optimal control theory' (Prentice-Hall, New Jersey, 1970), pp. 123–227
5 ELSGOLC, L.E.: 'Calculus of variations' (Addison-Wesley, Mass., 1962)
6 ANDERSON, B.D.O.: 'An algebraic solution to the spectral factorization problem', *IEEE Trans.*, 1967, AC-12, pp. 410–414
7 BLACK, K.F., and DENMAN, E.D.: 'Spectral factorization of proper matrix polynomials', *Int. J. Electron.*, 1977, 42, pp. 569–579
8 KALMAN, R.E.: 'When is a linear control system optimal?' *J. Basic Eng.*, 1964, 86, pp. 51–60
9 CHANG, S.S.L.: 'Synthesis of optimal control systems' (McGraw-Hill, New York, 1961)
10 TYLER, J.S., and TUTEUR, F.B.: 'The use of a quadratic performance index to design multivariable control systems', *IEEE Trans.*, 1966, AC-11, pp. 84–92
11 MOLINARI, B.P.: 'The stable regulator problem and its inverse', *ibid.*, 1973, AC-18, pp. 454–459
12 ANDERSON, K.W., and SHANNON, G.F.: 'Computational methods for solving the inverse-suboptimal-control problem', *Proc. IEE*, 1975, 122, (3), pp. 321–324
13 MUELLER, G.S.: 'Linear model of 2-shaft turbojet and its properties', *ibid.*, 1971, 118, pp. 813–815
14 McMORRAN, P.D.: 'Design of gas-turbine controller using inverse Nyquist method', *ibid.*, 1970, 117, pp. 2050–2056
15 TIWARI, R.N., PURKAYASTHA, P., and TIWARI, S.N.: 'Synthesis of stable and optimal controllers for a 2-shaft gas turbine', *ibid.*, 1977, 124, (12), pp. 1243–1248
16 OGATA, K.: 'Modern control engineering' (Prentice-Hall, New Jersey, 1970), pp. 216–258

8 Appendix

Block linear transformation

Consider a class of completely controllable, linear, time-invariant, multi-input, multi-output system

$$\dot{x}_0(t) = A_0 x_0(t) + B_0 u(t) \tag{36a}$$

$$y(t) = C_0 x_0(t) \tag{36b}$$

where $A_0 \in R^{n \times n}$, $B_0 \in R^{n \times m}$, $C_0 \in R^{l \times n}$, $x_0(t) \in R^{n \times 1}$, $y(t) \in R^{l \times 1}$, $u(t) \in R^{m \times 1}$. Assume that $l, m < n$ and $n/m = k$ (an integer) and define $r = n - m$. By a linear transformation

$$x_0(t) = T_1 z_1(t) \tag{37}$$

We wish to construct a state equation in the controllable block companion form

$$\dot{z}_1(t) = A_1 z_1(t) + B_1 u(t) \tag{38a}$$

$$y(t) = C_1 z_1(t) \tag{38b}$$

where

$$A_1 = T_1^{-1} A_0 T_1 = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] = \left[\begin{array}{ccccc|c} O_m & I_m & O_m & O_m & \cdots & O_m \\ O_m & O_m & I_m & O_m & \cdots & O_m \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ O_m & O_m & O_m & O_m & \cdots & I_m \\ \hline -D_1 & -D_2 & -D_3 & -D_4 & \cdots & -D_k \end{array}\right] \tag{38c}$$

$$B_1 = T_1^{-1} B_0 = \begin{bmatrix} O_{r \times m} \\ I_{m \times m} \end{bmatrix}, C_1 = C_0 T_1 = [N_1, N_2, \ldots, N_k], \tag{38d}$$

$A_{11} \in R^{r \times r}$, $A_{12} \in R^{r \times m}$, $A_{21} \in R^{m \times r}$, and $A_{22} \in R^{m \times m}$.
The constant matrices $D_i \in R^{m \times m}$ and $N_i \in R^{l \times m}$ are called block elements and the matrix $I_m = I_{m \times m} \in R^{m \times m}$ is an identity matrix. The matrices $O_m = O_{m \times m} \in R^{m \times m}$ and $O_{r \times m} \in R^{r \times m}$ are null matrices, respectively. The corresponding matrix transfer function of eqn. 38 can be directly formulated as

$$\begin{aligned} Y(s) &= [N_1 + N_2 s + \ldots + N_k s^{k-1}][D_1 + D_2 s + \ldots \\ &\quad + D_k s^{k-1} + I_m s^k]^{-1} U(s) \\ &= N(s) D^{-1}(s) U(s) = T_r(s) U(s) \end{aligned} \tag{39}$$

where $U(s)$ and $Y(s)$ are Laplace transforms of $u(t)$ and $y(t)$, and $T_r(s)$ is a matrix transfer function.

The objective is to derive the linear-transformation matrix T_1 in eqn. 37. Because T_1 transforms a state equation in eqn. 36 to a block companion form in eqn. 38, T_1 is called as a block linear transformation. We further assume that the matrix B_0 in eqn. 36 can be partitioned into the form of $\begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix}$ where $B_{21} \in R^{m \times m}$ is a non-singular matrix. This can be accomplished by rearranging the sequence of the elements in the state vector $x_0(t)$ in eqn. 36. By applying the first linear transformation

$$x_0(t) = K_1 \bar{x}_1(t) \tag{40}$$

where

$$K_1 = \begin{bmatrix} I_{r \times r} & B_{11} \\ O_{m \times r} & B_{21} \end{bmatrix} \quad \text{and} \quad K_1^{-1} = \begin{bmatrix} I_{r \times r} & -B_{11} B_{21}^{-1} \\ O_{m \times r} & B_{21}^{-1} \end{bmatrix}$$

we have

$$\dot{\bar{x}}_1(t) = \bar{A}_1 \bar{x}_1(t) + \bar{B}_1 u(t) \tag{41a}$$

$$y(t) = \bar{C}_1 \bar{x}_1(t) \tag{41b}$$

where

$$\bar{A}_1 = K_1^{-1} A_0 K_1 = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}, \quad \bar{B}_1 = K_1^{-1} B_0 = \begin{bmatrix} O_{r \times m} \\ I_{m \times m} \end{bmatrix},$$

$\bar{C}_1 = C_0 K_1$, $\bar{A}_{11} \in R^{r \times r}$, $\bar{A}_{12} \in R^{r \times m}$, $\bar{A}_{21} \in R^{m \times r}$, and $\bar{A}_{22} \in R^{m \times m}$.

To obtain the required state equation in eqn. 38, we perform the second linear transformation

$$\bar{x}_1(t) = K_2 z_1(t) \tag{42a}$$

where

$$K_2 = \begin{bmatrix} Q_1^{-1} & O_{r \times m} \\ -Q_2 Q_1^{-1} & I_{m \times m} \end{bmatrix}, \quad K_2^{-1} = \begin{bmatrix} Q_1 & O_{r \times m} \\ Q_2 & I_{m \times m} \end{bmatrix} \tag{42b}$$

and

$$[Q_1^T \,\vdots\, Q_2^T] = [q_1, \ldots, q_r \,\vdots\, q_{r+1}, \ldots, q_n]. \tag{42c}$$

T designates the transpose of the matrix. The unknown matrices $Q_1^T \in R^{r \times r}$ (with r column vectors q_i) and $Q_2^T \in R^{r \times m}$ (with m column vectors q_i) can be evaluated as follows.

From eqn. 42*a*, 41*a* and 38*c* we have the matrix equation

$$K_2^{-1} \bar{A}_1 = A_0 K_2^{-1} \tag{43a}$$

or

$$\begin{bmatrix} Q_1 & O_{r \times m} \\ Q_2 & I_{m \times m} \end{bmatrix} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} Q_1 & O_{r \times m} \\ Q_2 & I_{m \times m} \end{bmatrix} \tag{43b}$$

Expanding eqn. 43*b* yields

$$\begin{aligned} Q_1 \bar{A}_{11} &= A_{11} Q_1 + A_{12} Q_2 \\ Q_1 \bar{A}_{12} &= A_{12} \end{aligned} \tag{43c}$$

and

$$\begin{aligned} Q_2 \bar{A}_{11} + \bar{A}_{21} &= A_{21} Q_1 + A_{22} Q_2 \\ Q_2 \bar{A}_{12} + \bar{A}_{22} &= A_{22} \end{aligned} \tag{43d}$$

Performing a transpose operation on eqn. 43*c* and substituting eqns. 38*c* and 42*c* into it, we have the following recursive formulas:

$$\bar{A}_{11}^T q_i = q_{m+i} \quad \text{for} \quad i = 1, 2, \ldots, r \tag{44a}$$

$$\bar{A}_{12}^T q_i = O_{m \times 1} \quad \text{for} \quad i = 1, 2, \ldots, r-m \tag{44b}$$

and

$$\bar{A}_{12}^T q_{r-m+i} = e^i \qquad \text{for} \quad i = 1, 2, \ldots, m \tag{44c}$$

where e^i is the $m \times 1$ unit column vector whose ith element is unity, and all other elements are zeros. Eqn. 44 can be further simplified as follows:

(i) If $k = 2$ then

$$q_i = (\bar{A}_{12}^T)^{-1} e^i \quad \text{for} \quad i = 1, 2, \ldots, m \tag{45a}$$

and

$$q_{m+i} = \bar{A}_{11}^T q_i \qquad \text{for} \quad i = 1, 2, \ldots, m \tag{45b}$$

(ii) If $k > 2$, then

$$q_i = \begin{bmatrix} \bar{A}_{12}^T & & & \\ \bar{A}_{12}^T & (\bar{A}_{11}^T)^1 & & \\ \vdots & & & \\ \bar{A}_{12}^T & (\bar{A}_{11}^T)^{k-3} & & \\ \bar{A}_{12}^T & (\bar{A}_{11}^T)^{k-2} & & \end{bmatrix}^{-1} \begin{bmatrix} O_{m \times 1} \\ O_{m \times 1} \\ \vdots \\ O_{m \times 1} \\ e^i \end{bmatrix} \quad \text{for } i = 1, 2, \ldots, m \tag{45c}$$

and

$$q_{jm+i} = \bar{A}_{11}^T q_{(j-1)m+i} \quad \text{for} \quad i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, k-1. \tag{45d}$$

When the square matrices in eqns. 45*a* and 45*c* are not singular, the q_i in eqn. 42 can be obtained. Note that the determination of q_i in eqn. 45 only involves one inversion of a matrix. Thus the transformation matrix T_1 in eqn. 37, which links the co-ordinates $x_0(t)$ in eqn. 36 and the required co-ordinates $z_1(t)$ in eqn. 38, is

$$x_0(t) = K_1 K_2 z_1(t) = T_1 z_1(t) \tag{46}$$

It is believed that the block linear transformation T_1 is new.

An illustrative example

Consider the dynamic equation of an actual gas-turbine system[13] which is completely controllable and observable.

$$\dot{x}_0(t) = A_0 x_0(t) + B_0 u(t)$$
$$y(t) = C_0 x_0(t) \tag{47}$$

where

$$A_0 = \begin{bmatrix} -1{\cdot}268 & -0{\cdot}04528 & 1{\cdot}498 & 951{\cdot}5 \\ 1{\cdot}002 & -1{\cdot}957 & 8{\cdot}52 & 1240 \\ 0 & 0 & -10 & 0 \\ 0 & 0 & 0 & -100 \end{bmatrix}$$

$$B_0 = \begin{bmatrix} B_{11} \\ \hline B_{21} \end{bmatrix} = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ \hline 10 & 0 \\ 0 & 100 \end{array}\right] \quad C_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$n = 4$, $l = m = 2$, $r = n - m = 2$, and $k = n/m = 2$. The block companion form in eqn. 38, the corresponding matrix transfer function, of this system are required.

Applying the linear transformation in eqn. 40 yields the state equation in eqn. 41

$$\dot{x}_1(t) = \bar{A}_1 x_1(t) + \bar{B}_1 u(t)$$
$$y(t) = \bar{C}_1 x_1(t) \tag{48}$$

where

$$\bar{A}_1 = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \hline \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} = \left[\begin{array}{cc|cc} -1{\cdot}268 & -0{\cdot}04528 & 14{\cdot}98 & 95150 \\ 1{\cdot}002 & -1{\cdot}957 & 85{\cdot}2 & 124000 \\ \hline 0 & 0 & -10 & 0 \\ 0 & 0 & 0 & -100 \end{array}\right]$$

$$\bar{B}_1 = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ \hline 1 & 0 \\ 0 & 1 \end{array}\right] \quad \bar{C}_1 = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right]$$

and

$$K_1 = \begin{bmatrix} I_{r \times r} \\ \hline O_{m \times r} \end{bmatrix} \begin{bmatrix} B_{11} \\ \hline B_{21} \end{bmatrix} = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 100 \end{array}\right]$$

Applying the recursive algorithm in eqn. 45*a*, we have

$$q_1 = (\bar{A}_{12}^T)^{-1} e^1 = \begin{bmatrix} 14{\cdot}98 & 85{\cdot}2 \\ 95150 & 124000 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -1{\cdot}98423 \times 10^{-2} \\ 1{\cdot}52258 \times 10^{-2} \end{bmatrix}$$

$$q_2 = (\bar{A}_{12}^T)^{-1} e^2 = \begin{bmatrix} 14{\cdot}98 & 85{\cdot}2 \\ 95150 & 124000 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1{\cdot}36336 \times 10^{-5} \\ -2{\cdot}39708 \times 10^{-6} \end{bmatrix}$$

and

$$q_3 = \bar{A}_{11}^T q_1 = \begin{bmatrix} -1{\cdot}268 & 1{\cdot}002 \\ -0{\cdot}04528 & -1{\cdot}957 \end{bmatrix} \begin{bmatrix} -1{\cdot}98423 \times 10^{-2} \\ 1{\cdot}52258 \times 10^{-2} \end{bmatrix}$$

$$= \begin{bmatrix} 4{\cdot}04164 \times 10^{-2} \\ -2{\cdot}88984 \times 10^{-2} \end{bmatrix}$$

$$q_4 = \bar{A}_{11}^T q_2 = \begin{bmatrix} -1{\cdot}268 & 1{\cdot}002 \\ -0{\cdot}04528 & -1{\cdot}957 \end{bmatrix} \begin{bmatrix} 1{\cdot}36336 \times 10^{-5} \\ -2{\cdot}39708 \times 10^{-6} \end{bmatrix}$$

$$= \begin{bmatrix} -1{\cdot}96893 \times 10^{-5} \\ 4{\cdot}073763 \times 10^{-6} \end{bmatrix} \tag{49}$$

The transformation matrix K_2 in eqn. 42*b* is

$$K_2^{-1} = \begin{bmatrix} Q_1 \\ \hline Q_2 \end{bmatrix} \begin{bmatrix} O_{r \times m} \\ \hline I_{m \times m} \end{bmatrix}$$

$$= \left[\begin{array}{cc|cc} -1{\cdot}98423 \times 10^{-2} & 1{\cdot}52258 \times 10^{-5} & 0 & 0 \\ 1{\cdot}36336 \times 10^{-5} & -2{\cdot}39708 \times 10^{-6} & 0 & 0 \\ \hline 4{\cdot}04164 \times 10^{-2} & -2{\cdot}88984 \times 10^{-2} & 1 & 0 \\ -1{\cdot}96893 \times 10^{-5} & 4{\cdot}073763 \times 10^{-6} & 0 & 1 \end{array}\right]$$

The block linear transformation T_1 in eqn. 46 is

$$x_0(t) = K_1 K_2 z_1(t) = T_1 z_1(t) \tag{50}$$

where

$$T_1 = \begin{bmatrix} 14{\cdot}98 & 95150 & 0 & 0 \\ 85{\cdot}2 & 124000 & 0 & 0 \\ 18{\cdot}5671 & -2622{\cdot}1 & 10 & 0 \\ -5{\cdot}21389 \times 10^{-3} & 136{\cdot}829 & 0 & 100 \end{bmatrix}$$

The required block companion form in eqn. 38 is

$$\dot{z}_1(t) = A_1 z_1(t) + B_1 u(t) \tag{51}$$
$$y(t) = C_1 z_1(t)$$

where

$$A_1 = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline -18{\cdot}5671 & 2622{\cdot}1 & -11{\cdot}8567 & 262{\cdot}21 \\ 5{\cdot}214 \times 10^{-3} & -136{\cdot}83 & 0 & -101{\cdot}368 \end{array}\right]$$

$$B_1 = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ \hline 1 & 0 \\ 0 & 1 \end{array}\right] \quad C_1 = \left[\begin{array}{cc|cc} 14{\cdot}98 & 95150 & 0 & 0 \\ 85{\cdot}2 & 124000 & 0 & 0 \end{array}\right]$$

The corresponding matrix transfer function in eqn. 39 is

$$Y(s) = [N_1 + N_2 s]\,[D_1 + D_2 s + I_2 s^2]^{-1} U(s) \qquad (52)$$

where

$$N_1 = \begin{bmatrix} 14{\cdot}98 & 95150 \\ 85{\cdot}2 & 124000 \end{bmatrix} \quad N_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} 18{\cdot}5671 & -2622{\cdot}1 \\ -5{\cdot}214 \times 10^{-3} & 136{\cdot}83 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 11{\cdot}8567 & -262{\cdot}21 \\ 0 & 101{\cdot}368 \end{bmatrix}, \quad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

INT. J. SYSTEMS SCI., 1979, VOL. 10, NO. 12, 1415–1426

A geometric series approach to modelling discrete-time state equations from continuous-time state equations

L. S. SHIEH†, R. E. YATES‡, J. P. LEONARD‡
and J. M. NAVARRO§

A geometric series approach for approximating the state transition matrix of a continuous-time state equation by a discrete-time state transition matrix is developed in this paper. A discrete-time state equation is constructed for approximation of the continuous-time state equation. The discrete-time state transition matrix is modified to derive a generalized approximate numerical differentiator. Also, it is shown that several commonly used conversion procedures are special cases of this method.

1. Introduction

The accurate description of many practical systems often require high order continuous-time state equations. As a result, the simulation, realization and design of these high order systems are difficult. However, if the high order modelled continuous-time system can be represented by a discrete-time state equation, the analysis and implementation can be more easily accomplished by use of either a digital computer or a microprocessor. There exist several methods for converting the continuous-time state equations to the discrete-time state equations. One method involves analytical determination of the continuous-time state transition matrix of the system and converting it to the discrete-time state transition matrix for obtaining an exact discrete-time state equation. However, for a large system, this method is impractical. Another popular method (Bosley 1977) involves determination of an approximate transition matrix by truncating an infinite series that represents the exact state transition matrix. The truncating error of this approach depends heavily on the number of terms and the sampling period used. Other methods have been based on Tustin model (Cadzow 1973), Walsh function (Chen and Hsiao 1975) and Block pulse function (Shieh *et al.* 1978). These methods allow representation of a continuous-time state equation by an approximate discrete-time state equation derived from the trapezoid rule. In this paper, it will be shown that the approximate model (Cadzow 1973, Chen and Hsiao 1975, Shieh *et al.* 1978) so obtained is a special case of the models proposed.

A geometric series (Sherwood and Taylor 1952) approach is presented in this paper to approximate the discrete-time state transition matrix. Then the approximate discrete-time state transition matrix is used to construct an

Received 19 December 1978.

† Department of Electrical Engineering, University of Houston, Houston, Texas 77004, U.S.A.

‡ Guidance and Control Directorate, U.S. Army Missile Research and Development Command, Redstone Arsenal, Alabama 35809, U.S.A.

§ Departamento de Ingenieria Electronica, Instituto Universitario Politecnico, Barquisimeto, Venezuela.

0020-7721/79/1012 1415 $02·00

approximate discrete-time state equation. Also, the approximate discrete-time state transition matrix is modified to derive generalized approximate numerical differentiators (Jury 1964, Tou 1959).

2. Derivation of approximate discrete-time state equation

Consider the system represented by the continuous-time state equation

$$\left.\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t)\\ x(0) &= \alpha\end{aligned}\right\} \tag{1}$$

where $u(t)$ is the continuous-time input function.

For practical consideration (Jury 1964) we are interested in staircase inputs, or

$$u(t) = u(kT) \triangleq u(k)$$

for $k = 0, 1, 2, \ldots$, and $T =$ a sampling period and $kT \leqslant t < (k+1)T$.

The solution of eqn. (1) is

$$x(k+1) = \Phi(T)x(k) + Lu(k) \tag{2 a}$$

or

$$x(k) = \Phi(T)^k x(0) + \sum_{j=0}^{k-1} \Phi(k-j-1)Lu(j) \tag{2 b}$$

where

$$x(kT) \triangleq x(k)$$

$$x(kT+T) \triangleq x(k+1)$$

$$\Phi(kT-jT-T) \triangleq \Phi(k-j-1)$$

$\Phi(T)^k =$ the continuous-time state transition matrix

$$= [\exp(AT)]^k = \left[\sum_{j=0}^{\infty} \frac{1}{j!}(AT)^j\right]^k \tag{2 c}$$

$$L = \int_0^T \exp(A\alpha)B\,d\alpha = T\sum_{j=0}^{\infty} \frac{1}{(j+1)!}(AT)^j B$$

$$= [\exp(AT) - I]A^{-1}B = [\Phi(T) - I]A^{-1}B \tag{2 d}$$

where

$$\alpha = T - \lambda$$

For ease in implementation and manipulation we are interested in representing a continuous-time state equation by a discrete-time state equation :

$$\left.\begin{aligned}x_0^*(k+1) &= Dx_0^*(k) + Eu(k)\\ x_0^*(0) &= x(0)\end{aligned}\right\} \tag{3 a}$$

where

$$x_0^*(kT) \triangleq x_0^*(k) \cong x(kT)$$

$$x_0^*(kT+T) \triangleq x_0^*(k+1) \cong \dot{x}(kT)$$

The solution of eqn. (3 *a*) is

$$x_0^*(k) = D^k x(0) + \sum_{j=0}^{k-1} D^{k-j-1} E u(j) \tag{3 b}$$

where

$D^k = [\Phi_0^*(T)]^k$ is the discrete-time state transition matrix

$$\cong [\Phi(T)]^k = [\exp(AT)]^k \tag{3 c}$$

$$E = [D - I]A^{-1}B \cong [\phi(T) - I]A^{-1}B \tag{3 d}$$

A natural question is how accurately can one approximate $\Phi(T)$ by D? One popular method is to approximate $\Phi(T)$ in eqn. (2 *c*) by truncating the infinite series; i.e.

$$\Phi(T) = I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \frac{1}{4!}(AT)^4 + \ldots \tag{4 a}$$

$$\cong I + AT \tag{4 b}$$

$$\cong I + AT + \frac{1}{2!}(AT)^2 \tag{4 c}$$

$$\cong I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 \tag{4 d}$$

$$\cong I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \frac{1}{4!}(AT)^4 \tag{4 e}$$

$$\cong \ldots$$

If a sufficiently large number of terms in eqn. (4 *a*) is used, a satisfactory approximation may be obtained. However, the approximation error depends heavily on the number of terms and the sampling period used.

This paper introduces a geometric series which accurately approximates the infinite series in eqn. (4 *a*). Now, rewriting eqn. (4 *a*)

$$\Phi(T) = \exp(AT)$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \ldots + \frac{1}{j!}(AT)^j + \frac{1}{(j+1)!}(AT)^{j+1}$$

$$+ \frac{1}{(j+2)!}(AT)^{j+2} + \frac{1}{(j+3)!}(AT)^{j+3} + \ldots$$

$$+ \frac{1}{(j+n)!}(AT)^{j+n} + \ldots \tag{5}$$

Keeping the first $(j+1)$ important terms in the infinite series of eqn. (5) and approximating the rest of the terms in the equation by a geometric series with a weighting factor $1/(j^n)(j!)$ for the term $(AT)^{j+n}$ (rather than $1/(j+n)! = 1/(j+n)(j+n-1)\ldots(j+1)(j!)$ for the same term) in eqn. (5), we have an

accurate approximate model; i.e. the approximate model of $\Phi_0^*(T)$ using the proposed geometric series is

$$\Phi_0^*(T) = I + AT + \frac{1}{2!}(AT)^2 + \ldots + \frac{1}{j!}(AT)^j + \frac{1}{(j)(j!)}(AT)^{j-1} + \frac{1}{(j^2)(j!)}(AT)^{j+2} + \frac{1}{(j^3)(j!)}(AT)^{j+3} + \ldots + \frac{1}{(j^n)(j!)}(AT)^{j+n} + \ldots \quad (6\,a)$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \ldots + \frac{1}{j!}(AT)^j \left[I + \frac{1}{j}AT + \frac{1}{j^2}(AT)^2 + \frac{1}{j^3}(AT)^3 + \ldots + \frac{1}{j^n}(AT)^n + \ldots \right] \quad (6\,b)$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \ldots + \frac{1}{j!}(AT)^j \left[I - \frac{1}{j}AT \right]^{-1} \quad (6\,c)$$

$$= \left(I - \frac{1}{j}AT \right)^{-1} \left[I + \sum_{i=1}^{j-1} \frac{j-i}{(j)(i!)}(AT)^i \right] \quad \text{for } T < j/\|A\| \quad (6\,d)$$

$$\triangleq D_j \quad \text{for } j = 1, 2, 3, \ldots \quad (6\,e)$$

where $\|A\|$ is a matrix norm.

Note that the infinite series in the brackets of eqn. (6 *b*), or the term $[I-(1/j)AT]^{-1}$ in eqn. (6 *c*), is a geometric series. The subscript of D in eqn. (6 *e*) indicates the value of the weighting factor j to be used in the infinite series.

For example, when $j=2$, eqn. (5) can be approximated using the method of eqn. (6) by

$$\Phi(T) = I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \frac{1}{4!}(AT)^4 + \frac{1}{5!}(AT)^5 + \ldots \quad (7\,a)$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \sum_{j=3}^{\infty} \frac{1}{j!}(AT)^j \quad (7\,b)$$

$$\cong I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{(2)(2!)}(AT)^3 + \frac{1}{(2^2)(2!)}(AT)^4 + \frac{1}{(2^3)(2!)}(AT)^5 + \ldots \quad (7\,c)$$

$$= I + AT + \tfrac{1}{2}(AT)^2 + \frac{1}{2^2}(AT)^3 + \frac{1}{2^3}(AT)^4 + \frac{1}{2^4}(AT)^5 + \ldots \quad (7\,d)$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \sum_{j=3}^{\infty} \frac{1}{2^{j-1}}(AT)^j \quad (7\,e)$$

$$= I + AT + \tfrac{1}{2}(AT)^2 \left[I + \tfrac{1}{2}AT + \frac{1}{2^2}(AT)^2 + \frac{1}{2^3}(AT)^3 + \ldots \right] \quad (7\,f)$$

$$= I + AT + \tfrac{1}{2}(AT)^2(I - \tfrac{1}{2}AT)^{-1} \tag{7 g}$$

$$= (I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT) \quad \text{for } T < 2/\|A\| \tag{7 h}$$

$$\triangleq D_2 \tag{7 i}$$

The infinite series in the bracket of eqn. (7 *f*) is the geometric series. Comparing eqns. (4 *c*), (7 *e*) with (7 *b*) we note that the first three terms in all three equations are identical, but the other terms differ by their weighting factors. That is, zero in eqn. (4 *c*), and $1/(2^{j-1})$ in eqn. (7 *e*) and $1/j!$ in eqn. (7 *b*). Both eqns. (7 *b*) and (7 *e*) are infinite series. Clearly eqn. (7 *e*) is more accurate than that of eqn. (4 *c*).

In a like manner, when $j = 3$, we have

$$\Phi(T) = I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \sum_{j=4}^{\infty} \frac{1}{j!}(AT)^j \tag{8 a}$$

$$\cong I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 + \sum_{j=4}^{\infty} \frac{1}{(3^{j-3})(3!)}(AT)^j \tag{8 b}$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 \left[I + \tfrac{1}{3}AT + \frac{1}{3^2}(AT)^2 + \ldots \right] \tag{8 c}$$

$$= I + AT + \frac{1}{2!}(AT)^2 + \frac{1}{3!}(AT)^3 (I - \tfrac{1}{3}AT)^{-1} \tag{8 d}$$

$$= (I - \tfrac{1}{3}AT)^{-1}[I + \tfrac{2}{3}AT + \tfrac{1}{6}(AT)^2] \quad \text{for } T < 3/\|A\| \tag{8 e}$$

$$= \Phi^*(T) \tag{8 f}$$

$$\triangleq D_3 \tag{8 g}$$

Comparing eqns. (4 *d*), (8 *b*) and (8 *a*) we note that the first four terms in all three equations are identical, while the others differ by their weighting factors: zero in eqn. (4 *d*); $1/(3^{j-3})(3!)$ in eqn. (8 *b*); and $1/j!$ in eqn. (8 *a*). Also, comparing eqns. (7 *e*), (8 *b*) and (8 *a*), we observe that eqn. (8 *b*) consists of one more important term, $(1/3!)(AT)^3$, than that of eqn. (7 *e*) and the approximation of $(1/(3^{j-3})(3!))(AT)^j$ in eqn. (8 *b*) to $(1/j!)(AT)^j$ in eqn. (8 *a*) for $j = 4, 5, \ldots$ is better than that of $(1/2^{j-1})(AT)^j$ in eqn. (7 *e*) to $(1/j!)(AT)^j$ in eqn. (8 *a*) for $j = 4, 5, \ldots$. Therefore, the approximate model in eqn. (8 *e*) is more accurate than the model in eqn. (7 *h*). The approximate models of D_j, for $j = 1, 2, \ldots$, can be obtained from eqn. (6 *d*) and are as follows:

$$D \cong (I - AT)^{-1} \triangleq D_1$$

$$\cong (I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT) \triangleq D_2$$

$$\cong (I - \tfrac{1}{3}AT)^{-1}[I + \tfrac{2}{3}AT + \tfrac{1}{6}(AT)^2] \triangleq D_3$$

$$\cong (I - \tfrac{1}{4}AT)^{-1}[I + \tfrac{3}{4}AT + \tfrac{1}{4}(AT)^2 + \tfrac{1}{24}(AT)^3] \triangleq D^4$$

$$\cong (I - \tfrac{1}{5}AT)^{-1}[I + \tfrac{4}{5}AT + \tfrac{3}{10}(AT)^2 + \tfrac{1}{15}(AT)^3 + \tfrac{1}{120}(AT)^4] \triangleq D^5$$

$$\cong \ldots \tag{9 a}$$

D_2 in eqn. (9 *a*) is a commonly used model (Cadzow 1973, Chen and Hsiao 1975, Shieh *et al.* 1978, Jury 1964, Tou 1959). In general, the larger j used in eqn. (6) provides a more accurate model.

Substituting the approximate models D_j in eqn. (9 *a*) into eqn. (3 *d*) yields the input matrix E, or

$$\begin{aligned} E &= (D_j - I)A^{-1}B \\ &\cong T(I-AT)^{-1}B \triangleq E_1 \\ &\cong T(I-\tfrac{1}{2}AT)^{-1}B \triangleq E_2 \\ &\cong T(I-\tfrac{1}{3}AT)^{-1}(I+\tfrac{1}{6}AT)B \triangleq E_3 \\ &\cong T(I-\tfrac{1}{4}AT)^{-1}[I+\tfrac{1}{4}AT+\tfrac{1}{24}(AT)^2]B \triangleq E_4 \\ &\cong T(I-\tfrac{1}{5}AT)^{-1}[I+\tfrac{3}{10}AT+\tfrac{1}{15}(AT)^2+\tfrac{1}{120}(AT)^3]B \triangleq E_5 \\ &\cong \ldots \end{aligned} \tag{9 b}$$

An alternate form of eqn. (3) is

$$x^*(k+1) = Gx^*(k) + Hu(k) \tag{10}$$

$$x^*(0) = x(0)$$

The G can be obtained by modifying the $\exp(AT)$ as follows:

$$G \cong \exp(AT) = [\exp(-\tfrac{1}{2}AT)]^{-1} \exp(\tfrac{1}{2}AT) \tag{11}$$

An approximation of $\exp(\frac{1}{2}AT)$ and $\exp(-\frac{1}{2}AT)$ can be obtained from eqn. (6) by replacing T in eqn. (6 *d*) by $\frac{1}{2}T$ and $-\frac{1}{2}T$, respectively. Finally, we have

$$G \cong [\exp(-\tfrac{1}{2}AT)]^{-1} \exp(\tfrac{1}{2}AT) \tag{12 a}$$

$$\cong \left(\left[I+\frac{1}{2j}AT\right]^{-1}\left[I+\sum_{i=1}^{j-1}\frac{(-1)^i(j-i)}{(2^i)(j)(i!)}(AT)^i\right]\right)^{-1} \times \left(\left[I-\frac{1}{2j}AT\right]^{-1}\left[I+\sum_{i=1}^{j-1}\frac{(j-i)}{(2^i)(j)(i!)}(AT)^i\right]\right) \tag{12 b}$$

$$\triangleq G_j \quad \text{for } j = 1, 2, 3, \ldots \tag{12 c}$$

$$\triangleq Q_j^{-1} P_j \quad \text{for } j = 1, 2, 3, \ldots \tag{12 d}$$

where

$$Q_j = \left[I-\frac{1}{2j}AT\right]\left[I+\sum_{i=1}^{j-1}\frac{(-1)^i(j-i)}{(2^i)(j)(i!)}(AT)^i\right] \tag{12 e}$$

$$P_j = \left[I+\frac{1}{2j}AT\right]\left[I+\sum_{i=1}^{j-1}\frac{(j-i)}{(2^i)(j)(i!)}(AT)^i\right] \tag{12 f}$$

The approximate system matrices G_j for $j = 1, \ldots, 4$ are

$$G_1 = [I-\tfrac{1}{2}AT]^{-1}[I+\tfrac{1}{2}AT] = Q_1^{-1} P_1 \tag{13 a}$$

$$G_2 = [I-\tfrac{1}{2}AT+\tfrac{1}{12}(AT)^2]^{-1}[I+\tfrac{1}{2}AT+\tfrac{1}{12}(AT)^2] = Q_2^{-1} P_2 \tag{13 b}$$

$$G_3=[I-\tfrac{1}{2}AT+\tfrac{7}{72}(AT)^2-\tfrac{1}{144}(AT)^3]^{-1}$$
$$\times[I+\tfrac{1}{2}AT+\tfrac{7}{72}(AT)^2+\tfrac{1}{144}(AT)^3]=Q_3^{-1}\,P_3 \qquad (13\,c)$$

$$G_4=[I-\tfrac{1}{2}AT+\tfrac{7}{84}(AT)^2-\tfrac{5}{384}(AT)^3+\tfrac{1}{1536}(AT)^4]^{-1}$$
$$\times[I+\tfrac{1}{2}AT+\tfrac{7}{84}(AT)^2+\tfrac{5}{384}(AT)^3+\tfrac{1}{1536}(AT)^4]=Q_4^{-1}\,P_4 \qquad (13\,d)$$

The approximate input matrices H_j for $j=1,\ldots,4$ are

$$H_j=[G_j-I]A^{-1}B \qquad (14\,a)$$

or

$$H_1=T[I-\tfrac{1}{2}AT]^{-1}B \qquad (14\,b)$$

$$H_2=T[I-\tfrac{1}{2}AT+\tfrac{1}{16}(AT)^2]^{-1}B \qquad (14\,c)$$

$$H_3=T[I-\tfrac{1}{2}AT+\tfrac{7}{72}(AT)^2-\tfrac{1}{144}(AT)^3]^{-1}[I+\tfrac{1}{72}(AT)^2]B \qquad (14\,d)$$

$$H_4=T[I-\tfrac{1}{2}AT+\tfrac{7}{84}(AT)^2-\tfrac{5}{384}(AT)^3+\tfrac{1}{1536}(AT)^4]^{-1}$$
$$\times[I+\tfrac{5}{192}(AT)^2]B \qquad (14\,e)$$

Noting that the coefficients of Q_j and P_j in eqn. (13) are identical except for signs, we will derive a general equation for an approximate numerical differentiator.

3. Approximate numerical differentiator

When $u(t)$ in eqn. (1) is a given input function in the analytical form or in discrete form, the input function $u(t)$ is often approximated by a trapezoid rule

$$u^*(k)=\frac{u(k+1)+u(k)}{2} \qquad (15)$$

where

$$u^*(k)\triangleq u^*(kT)$$

The approximate discrete-time state equation is

$$x^*(k+1)=Gx^*(k)+Hu^*(k) \qquad (16\,a)$$
$$=Q_j^{-1}\,P_jx^*(k)+H_ju^*(k) \qquad (16\,b)$$

where G and H are shown in eqns. (12) and (14 *a*), respectively. For example, if $G=G_1$ in eqn. (13 *a*) and $H=H_1$ in eqn. (14 *b*) are used, we have

$$x^*(k+1)=(I-\tfrac{1}{2}AT)^{-1}(I+\tfrac{1}{2}AT)x^*(k)+T(I-\tfrac{1}{2}AT)^{-1}Bu^*(k) \qquad (17)$$

Equation (17) can be derived from eqn. (1) by using the Walsh function and the Block-pulse function approaches (Chen and Hsiao 1975, Shieh *et al.* 1978). Therefore, it is seen that the approximate discrete-time model obtained by the above approaches is a special case of the models presented in this paper. Since the coefficients of Q_j and P_j in eqn. (13) are identical except for signs, we can derive a generalized approximate numerical differentiator as follows.

Taking the z transform of eqn. (1) when $u(t)=$continuous input functions yields

$$Z[\dot{x}(t)]=AZ[x(t)]+BZ[u(t)] \qquad (18\,a)$$

or

$$Z[\dot{x}(kT)] = AZ[x(kT)] + BZ[u(kT)] \tag{18 b}$$

$$Z[\dot{x}(t)] = Z[\dot{x}(kT)]$$

$$= zx(z) - zx(0) = Ax(z) + Bu(z) \tag{18 c}$$

Also, taking the z transform of eqn. (16) for $j=1$ we have

$$Z[x^*(k+1)] = Q_1^{-1} P_1 Z[x^*(k)] + H_1 Z[u^*(k)]$$

or

$$zx^*(z) - zx^*(0) = (I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT)x^*(z)$$

$$+ T(I - \tfrac{1}{2}AT)^{-1}B \cdot \frac{(z+1)}{2} u(z) \tag{18 d}$$

Rearranging eqn. (18 d) yields

$$\frac{2}{T}\frac{(z-1)}{(z+1)} x^*(z) - \frac{2}{T}\frac{z}{(z+1)} (I - \tfrac{1}{2}AT)x(0) = Ax^*(z) + Bu(z) \tag{19}$$

Comparing eqns. (18 c) and (19) we have

$$Z[\dot{x}(t)] \cong \frac{2}{T}\frac{(z-1)}{(z+1)} x^*(z) - \frac{2}{T}\frac{z}{(z+1)} (I - \tfrac{1}{2}AT)x(0) \tag{20}$$

Equation (20) is the approximate numerical differentiator which is often used to determine the inverse Laplace transform of a continuous-time state equation (Jury 1964, Tou 1959). The general representation of the approximate numerical differentiator, based on the approximate models presented in this paper, is

$$Z[\dot{x}(t)] \cong \frac{AT}{2} (P_j - Q_j)^{-1}(P_j + Q_j) \frac{2}{T}\frac{(z-1)}{(z+1)} x^*(z)$$

$$- AT(P_j - Q_j)^{-1} Q_j \frac{2}{T}\frac{z}{(z+1)} x(0)$$

$$= A(P_j - Q_j)^{-1}(P_j + Q_j) \frac{(z-1)}{(z+1)} x^*(z)$$

$$- 2A(P_j - Q_j)^{-1} Q_j \frac{z}{(z+1)} x(0) \tag{21}$$

where P_j and Q_j are shown in eqn. (12).

For example, if $G = G_2 = Q_2^{-1} P_2$ and $H = H_2$ in eqns. (13) and (14) are used, we have

$$Z[\dot{x}(t)] \cong [I + \tfrac{1}{16}(AT)^2] \frac{2}{T}\frac{(z-1)}{(z+1)} x^*(z)$$

$$- [I - \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2] \frac{2}{T}\frac{z}{(z+1)} x(0) \tag{22}$$

4. Illustrative example

Consider an unstable continuous state equation :

$$\dot{x}(t) = Ax(t) + Bu(t)$$
$$x(0) = x_0 \tag{23}$$

where

$$A = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}, \quad x(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

and $u(t)$ is the unit-step functions.

The discrete responses, $x(kT)$ at $k = 0, 1, \ldots, 4$ and

$$T = \tfrac{1}{4} < 2/(|3| + |-4|) = \tfrac{2}{7},$$

using the approximate numerical differentiators for $j = 1$ and 2 in eqn. (21) are required.

When $j = 1$ in eqn. (12) we have

$$\left.\begin{aligned} Q_1 &= I - \tfrac{1}{2}AT \\ P_1 &= I + \tfrac{1}{2}AT \end{aligned}\right\} \tag{24}$$

The corresponding approximate numerical differentiator can be obtained from eqns. (21) and (24)

$$\begin{aligned} Z[\dot{x}(t)] &\cong A(P_1 - Q_1)^{-1}(P_1 + Q_1)\frac{(z-1)}{(z+1)}x^*(z) \\ &\quad - 2A(P_1 - Q_1)^{-1}Q_1\frac{z}{(z+1)}x(0) \\ &= \frac{2}{T}\frac{(z-1)}{(z+1)}x^*(z) - \frac{2}{T}(I - \tfrac{1}{2}AT)\frac{z}{(z+1)}x(0) \end{aligned} \tag{25}$$

Taking the z transform of eqn. (23) and substituting eqn. (25) into it we have

$$\begin{aligned} zx^*(z) - zx(0) &= (I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT)x^*(z) \\ &\quad + T(I - \tfrac{1}{2}AT)^{-1}B\frac{(z+1)}{2}u(z) \end{aligned} \tag{26}$$

The corresponding discrete state equation is

$$\begin{aligned} x^*(k+1) &= (I - \tfrac{1}{2}AT)^{-1}(I + \tfrac{1}{2}AT)x^*(k) + T(I - \tfrac{1}{2}AT)^{-1}Bu^*(k) \\ &= G_1x^*(k) + H_1u^*(k) \end{aligned} \tag{27}$$

where

$$G_1 = \begin{bmatrix} 1{\cdot}461538 & 0{\cdot}410256 \\ 0{\cdot}615384 & 0{\cdot}435897 \end{bmatrix} \quad \text{and} \quad H_1 = \begin{bmatrix} 0{\cdot}666666 & 0{\cdot}051282 \\ 0{\cdot}333333 & 0{\cdot}179487 \end{bmatrix}$$

and

$$u^*(k) = \frac{u(k+1) + u(k)}{2}$$

Equation (27) is a commonly used model (Shieh *et al.* 1978, Jury 1964, Tou 1959). When $j=2$ in eqn. (12) we have

$$\left.\begin{aligned} Q_2 &= I - \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2 \\ P_2 &= I + \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2 \end{aligned}\right\} \tag{28}$$

The approximate numerical differentiator in eqn. (21) becomes

$$Z[\dot{x}(t)] \cong [I + \tfrac{1}{16}(AT)^2]\,\frac{2}{T}\,\frac{(z-1)}{(z+1)}\,x^*(z) - [I - \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2]\cdot\frac{2}{T}\,\frac{z}{(z+1)}\,x(0) \tag{29}$$

The corresponding discrete state equation is

$$\begin{aligned} x^*(k+1) &= [I - \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2]^{-1}[I + \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2]x^*(k) \\ &\quad + T[I - \tfrac{1}{2}AT + \tfrac{1}{16}(AT)^2]^{-1}Bu^*(k) \\ &= G_2x^*(k) + H_2u^*(k) \end{aligned} \tag{30}$$

where

$$G_2 = \begin{bmatrix} 1{\cdot}456106 & 0{\cdot}393909 \\ 0{\cdot}590865 & 0{\cdot}471331 \end{bmatrix} \quad \text{and} \quad H_2 = \begin{bmatrix} 0{\cdot}653061 & 0{\cdot}051830 \\ 0{\cdot}326531 & 0{\cdot}171039 \end{bmatrix}$$

and

$$u^*(k) = \frac{u(k+1) + u(k)}{2}$$

The exact solution of eqn. (23) is

$$x_1(t) = \tfrac{16}{7}\exp(2t) - \tfrac{3}{35}\exp(-5t) - \tfrac{6}{5} \tag{31 a}$$

$$x_2(t) = \tfrac{8}{7}\exp(2t) + \tfrac{9}{35}\exp(-5t) - \tfrac{2}{5} \tag{31 b}$$

The responses at the sampling instants $k=0, 1, \ldots, 4$ of the exact solution and the two approximates are shown in Table 1 [state $x_1(kT)$] and Table 2 [state $x_2(kT)$]. From Tables 1 and 2 we note that better results are obtained with the improved model.

k	T	Exact solution	Approximate solution	
		Eqn. (31 *a*)	Eqn. (27)	Eqn. (30)
0	0·00	1	1	1
1	0·25	2·544	2·589	2·555
2	0·50	5·006	5·145	5·040
3	0·75	9·042	9·380	9·123
4	1·00	15·689	16·436	15·867

Table 1. Comparison of state $x_1(kT)$.

k	T	Exact solution	Approximate solution	
		Eqn. (31 *b*)	Eqn. (27)	Eqn. (30)
0	0·00	1	1	1
1	0·25	1·558	1·564	1·560
2	0·50	2·728	2·788	2·742
3	0·75	4·728	4·894	4·768
4	1·00	8·046	8·419	8·135

Table 2. Comparison of state $x_2(kT)$.

5. Conclusion

A geometric series approach has been presented for determination of a set of approximate discrete-time state equations from the continuous-time state equations. The approximate discrete-time models have been modified so that a generalized approximate numerical differentiator can be derived. It has also been shown that several commonly used conversion procedures are special cases of the method given in this paper.

We have also shown that the proposed geometric series approach approximates the exponential matrix infinite series in eqn. (5) by taking a finite number of dominant terms and an infinite number of the other terms of the matrix series expansion rather than taking a finite number of dominant terms only. However, the method requires a matrix inversion and the approximate models are valid only in the region where the geometric series is convergent, that is the sampling period T for the models in eqn. (6) is limited to $T \ll j/\|A\|$. These are the limitations of the proposed approach.

Despite these limitations, the proposed models can be effectively applied to perform the numerical integrations of stiff functions because the most commonly used model (i.e. G_1 in eqn. (13 *a*) and H_1 in eqn. (14 *b*) (Cadzow 1973, Chen and Hsiao 1975, Shieh 1978), which is the lowest order model proposed in this paper, has been successfully used for evaluating the responses of stiff functions (Chen and Hsiao 1975). Furthermore, a higher order model that uses a larger weighting factor j makes possible the use of a larger sampling period T. This observation can be verified from the fact that T is proportional to j (i.e. $T < j/\|A\|$ in eqn. (6)). This result will greatly increase the flexibility in determining the common sampling period among various subsystems of a large sampled-data control system.

Acknowledgments

This work was supported in part by the U.S. Army Missile Research and Development Command, DAAK 40-79-C-0061 and the U.S. Army Research Office, DAAG 29-79-C-0178.

References

Bosley, J. T., 1977, Final Report. DAAK40-77-C-0048 TGT-001.

Cadzow, J. A., 1973, *Discrete-Time Systems* (Englewood Cliffs, N.J.: Prentice-Hall), pp. 236–244.

Chen, C. F., and Hsiao, C. H., 1975, *Int. J. Systems Sci.*, **6**, 833.

JURY, E. I., 1964, *Theory and Application of the z-Transform Method* (New York : Wiley).
SHIEH, L. S., YEUNG, C. K., and MCINNIS, B. C., 1978, *Int. J. Control*, **28**, 383.
SHERWOOD, G. E. F., and TAYLOR, A. E., 1952, *Calculus* (New York : Prentice-Hall), pp. 371–392.
TOU, J. T., 1959, *Digital and Sampled-Data Control Systems* (New York : McGraw-Hill), pp. 208–209.

INT. J. SYSTEMS SCI., 1979, VOL. 10, NO. 10, 1097–1114

A method for modelling transfer functions using dominant frequency-response data and its applications

L. S. SHIEH†, M. DATTA-BARUA†, and R. E. YATES‡

This paper presents a fundamental method for modelling transfer functions using the basic performance specifications and frequency-response data at the dominant frequencies. A set of non-linear equations is constructed from the definitions of the basic performance specifications, the dominant frequency-response data and the unknown coefficients of a transfer function. A Newton–Raphson multidimensional method is applied to solve the non-linear equations. Four methods are given to construct approximate representations of the desired transfer functions for the estimation of good starting values to ensure rapid convergence of the numerical method. The applications of the proposed method are: (1) developing a standard model and/or a transfer function of a filter or a compensator using the specified dominant frequency-response data; (2) identifying the transfer function of a system from available experimental frequency-response data; and (3) reducing high-order transfer functions to low-order models using dominant frequency-response data.

1. Introduction

The nature of the transient response of a system is often characterized by a set of performance specifications in the time domain such as the settling time and the rising time. In the frequency domain, another set of performance specifications (Gibson and Rekasius 1961) is used to represent the characteristics of the system performance. The bandwidth and the phase margin are typical examples of the frequency domain specifications. In designing compensators and filters, and in predicting the nature of time response of a system, practicing engineers are often interested in the dominant poles. These can be converted to a damping ratio and a natural angular frequency specified in the complex plane. These specifications are often called the complex-domain specifications. The engineer is also interested in various error constants (for example, the velocity-error constant), which represent the characteristics of system performance in both time and frequency domains (Truxal 1955). The frequency-response data at the frequencies of the frequency-domain specification are considered as the dominant frequency-response data in this paper because these data characterize the nature of the system responses. For example, the phase margin (ϕ_m) of a system at the gain-crossover frequency (ω_c) is often used as a measure of additional phase lag required to bring the system to the verge of instability. Also, if the phase angle of the open-loop system at the ω_c is near $-180°$, then the response of the closed-loop system will be oscillatory.

Received 5 July 1978.

† Department of Electrical Engineering, University of Houston, Houston, Texas 77004, U.S.A.

‡ Guidance and Control Directorate, U.S. Army Missile Research and Development Command, Redstone Arsenal, Alabama 35809, U.S.A.

0020-7721/79/1010 1007 $02·00

In the design of a control system in the frequency domain, the specifications discussed above or the dominant frequency-response data are usually considered as design goals. Various frequency-domain or complex-domain approaches (Nyquist 1932, Evans 1953, Bode 1954, Thaler 1973) have been developed and widely applied in industry for compensator designs to achieve desired performance. The most popular design methods are those based on the Nyquist (1932) plot, the Bode (1954) design, and the root-locus method (Evans 1953, Thaler 1973). To improve the efficiency of the design methods, it is advantageous to have the design goals expressed as mathematical functions or transfer functions (defined as the standard models). Once standard models have been ascertained, the corresponding time-domain specifications and temporal responses can be determined from digital or analogue simulations of the standard models. Also, the frequency-response data of the desired compensator can be determined from Nyquist plots or Bode diagrams by comparing the frequency-response curves of the original and the desired response models. The required filters and compensators (Del Toro and Parker 1960, Thaler 1973) can then be easily determined.

Empirical rules or rules of the thumb that link the specifications in the time, frequency, and complex domains have been developed by Truxal (1955), Del Toro and Parker (1960), Axelby (1960), and Seshadri (1969) *et al.* From these results, it is observed that most time-domain specifications and complex-domain specifications can be approximately converted to frequency-domain specifications. Some of these frequency-domain specifications are phase margin (ϕ_m), maximum value of the closed-loop frequency response (M_p), gain-crossover frequency (ω_c), peak value frequency (ω_p), the bandwidth (ω_b), and velocity-error constant (K_v). Other important frequency-response data are: (1) the real part of the open-loop transfer function $G(j\omega)$ at the phase-crossover frequency (ω_π) which has been used to define the gain margin (G_m); (2) the real part and imaginary part of the closed-loop function ($T(s)$) and the open-loop function $G(s)$ at $s=j\omega \triangleq j\omega_0=j0$. The data at $\omega=0$ often indicate the final value and the type of the system. In a type 1 system, $I_m[G(j0)]$ has an infinite value, while Re $[G(j0)]$ has a finite value from which an asymptotic line (Del Toro and Parker 1960) can be drawn in a Nyquist plot; (3) the corner frequencies in the Bode plot of $G(j\omega)$ in the regions of $\omega=\omega_{c1}$ where $20 \log |G(j\omega_{c1})| = +15$ dB, and $\omega=\omega_{c2}$ where $20 \log |G(j\omega_{c2})| = -15$ dB. Chen (1957) has shown empirically that the open-loop poles and zeros of a system can be approximated by retaining the Bode plot in the regions of the ± 15 dB boundaries. Some dominant frequency-response data are indicated in Fig. 1.

Through use of the above dominant frequency-response data, a basic method is proposed in this paper for modelling various transfer functions. First, a set of simultaneous non-linear algebraic equations, based on basic definitions of the dominant frequency-response data and the unknown coefficients of a desired transfer function, is constructed. Then the Newton–Raphson method (Carnahan *et al.* 1969, IBM 1977) is used to solve the non-linear equations. However, as is well known, the Newton–Raphson method will often only converge for a small range of starting values; therefore, four methods are developed in this paper for estimating good starting values so that the numerical method (IBM 1977) will converge rapidly to the desired solution.

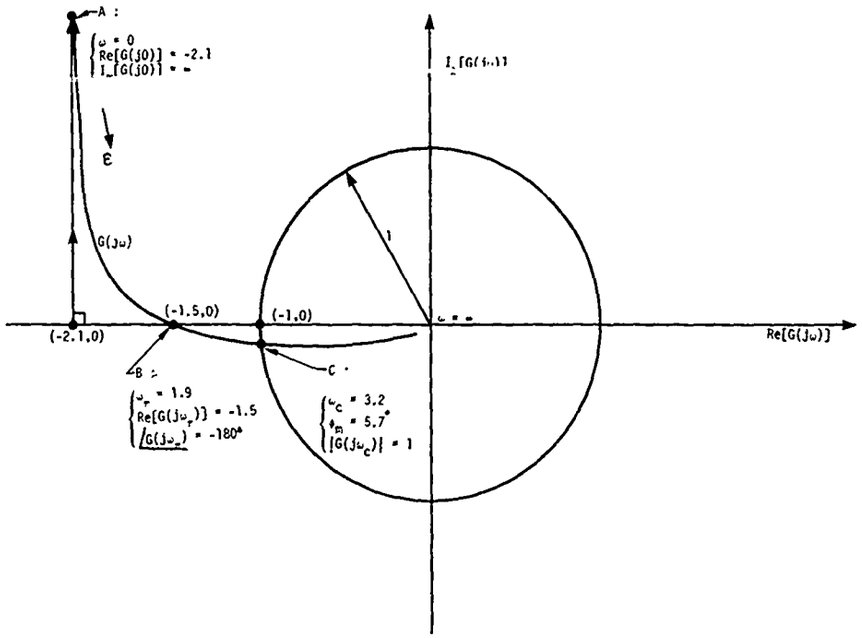


Figure 1. Nyquist plot of an open-loop system $G(s)$.

The applications of this method can be classified as follows.

(1) When the design goals are predescribed by the dominant frequency-response data, which may be obtained from the frequency-domain specifications (Gibson and Rekasius 1961) or equivalent ones (Truxal 1955, Del Toro and Parker 1960, Axelby 1960, Seshadzi *et al.* 1969), and a standard transfer function is desired, this is a *design problem.* Chen and Shieh (1970) and Wakeland (1976) have proposed analytical methods for the compensator fitting. However, their methods are limited to filters and compensators in which the unknown coefficients can be solved by a quadratic equation. The method of this paper overcomes this difficulty.

(2) The transfer function obtained in this paper is the function of the original system. When dominant frequency-response data can be obtained from experimental data of a practical system and the mathematical function of the system is desired, this is an *identification problem.*

(3) When the dominant frequency-response data are obtained from a given high-order transfer function and various low-order approximate models are required, this is the *model reduction problem.* The reduced models obtained in this paper have the same selected dominant frequency-response data as the original system. Thus, the design processes in the frequency domain can be greatly simplified.

2. Modelling non-linear equations

Given a transfer function $T(s)$ of a unity ratio feedback closed-loop system

$$T(s)=\frac{b_0+b_1s+b_2s^2+\ \ldots\ +b_ms^m}{a_0+a_1s+a_2s^2+\ \ldots\ +a_ns^n}=\frac{n(s)}{d(s)}=\frac{G(s)}{1+G(s)} \tag{1 a}$$

where $n(s)$ and $d(s)$ are the numerator and denominator polynomials, respectively, and a_i and b_i are constants. If the system is a type l system, the open-loop transfer function $G(s)$ is

$$G(s)=\frac{K(1+c_1s+c_2s^2+\ \ldots\ +c_ps^p)}{s^l(1+d_1s+d_2s^2+\ \ldots\ +d_qs^q)}=\frac{p(s)}{q(s)} \tag{1 b}$$

where $p(s)$ and $q(s)$ are the numerator and denominator polynomials. K, l, c_i, and d_i are constants. K is a velocity-error constant (K_v) if $l=1$.

The equations for dominant frequency-response data are :

(1) System type is determined from

$$G(j\omega_0)=\mathrm{Re}\,[G(j\omega_0)]+jI_m[G(j\omega_0)] \quad \text{at } \omega_0=0 \tag{2 a}$$

or

$$\left.\begin{aligned} G(j0)&=\mathrm{Re}\,[G(j0)]\\ T(j0)&=\frac{b_0}{a_0}\end{aligned}\right\} \quad \text{for a type 0 system} \tag{2 b}$$

$$\left.\begin{aligned} \mathrm{Re}\,[G(j0)]&=K(c_1-d_1)\\ I_m[G(j0)]&=\infty\\ T(j0)&=\frac{b_0}{a_0}=1\end{aligned}\right\} \quad \text{for a type 1 system} \tag{2 c}$$

(2) Phase margin gives

$$\phi_m=180^\circ+\angle G(j\omega_c) \tag{3 a}$$

where

$$|G(j\omega_c)|=1 \tag{3 b}$$

ω_c is the gain crossover-frequency.

(3) Gain margin yields

$$G_m=\left|\frac{1}{\mathrm{Re}\,[G(j\omega_\pi)]}\right| \tag{4 a}$$

where

$$\angle G(j\omega_\pi)=-180^\circ \tag{4 b}$$

ω_π is the phase crossover frequency.

(4) $M_p=|T(j\omega_p)|=$ maximum value of the closed-loop frequency response (5 a)

where

$$\left.\frac{d|T(j\omega)|}{d\omega}\right|_{\omega=\omega_p}=0 \tag{5 b}$$

ω_p is the peak value frequency.

(5) $$|T(j\omega_b)| = \frac{1}{\sqrt{2}} \tag{6}$$

where ω_b is the bandwidth.

(6) $$|G(j\omega_{c1})| = 5{\cdot}6 \tag{7 a}$$

or

$$20 \log |G(j\omega)| = +15 \text{ dB} \quad \text{at } \omega = \omega_{c1} \tag{7 b}$$

and

$$|G(j\omega_{c2})| = 0{\cdot}18 \tag{7 c}$$

or

$$20 \log |G(j\omega)| = -15 \text{ dB} \quad \text{at } \omega = \omega_{c2} \tag{7 d}$$

A set of non-linear equations can be formulated from the basic definitions of the assigned dominant frequency-response data in (2)–(7). The procedures can be illustrated by using the following example. The dominant frequency-response data in (2 *c*), (3), and (4) are shown in Fig. 1, which are marked as A, B, and C and given as follows :

(1) $\text{Re}\,[G(j\omega_0)] = -2{\cdot}1$ and $I_m[G(j\omega_0)] = \infty$ at $\omega_0 = 0$ rad/s

or $$T(j\omega_0) = 1 \quad \text{at } \omega_0 = 0 \text{ rad/s} \tag{8 a}$$

(2) $$\text{Re}\,[G(j\omega_n)] = -1{\cdot}5 \quad \text{at } \omega_n = 1{\cdot}9 \text{ rad/s} \tag{8 b}$$

(3) $$\angle G(j\omega_n) = -180^\circ \quad \text{at } \omega_n = 1{\cdot}9 \text{ rad/s} \tag{8 c}$$

(4) $$\phi_m = 180^\circ + \angle G(j\omega_c) = 5{\cdot}7^\circ \quad \text{at } \omega_c = 3{\cdot}2 \text{ rad/s} \tag{8 d}$$

(5) $$|G(j\omega_c)| = 1 \quad \text{at } \omega_c = 3{\cdot}2 \text{ rad/s} \tag{8 e}$$

Five conditions are given in (8). Therefore, various transfer functions with five unknown coefficients can be constructed. Assume that the desired transfer function $T_d(s)$ is

$$T_d(s) = \frac{b_0 + b_1 s + b_2 s^2}{a_0 + a_1 s + a_2 s^2 + a_3 s^3} \tag{9 a}$$

From the conditions in (8 *a*), it may be observed that the system is a type 1 system. Therefore $b_0 = a_0$. Also, to simplify the equation we let $a_3 = 1$. Thus, we have

$$T_d(s) = \frac{a_0 + b_1 s + b_2 s^2}{a_0 + a_1 s + a_2 s^2 + s^3} \tag{9 b}$$

The corresponding open-loop transfer function $G_d(s)$ is

$$G_d(s) = \frac{K(1 + c_1 s + c_2 s^2)}{s(1 + d_1 s + d_2 s^2)} \tag{10}$$

where

$$K = \frac{a_0}{a_1 - b_1}, \quad c_1 = \frac{b_1}{a_0}, \quad c_2 = \frac{b_2}{a_0}, \quad d_1 = \frac{a_2 - b_2}{a_1 - b_1} \quad \text{and} \quad b_2 = \frac{1}{a_1 - b_1}$$

Following the basic definitions and the assigned data in (8) yields a set of non-linear equations :

(1) The assignment in (8 *a*), or Re $[G(j0)] = -2{\cdot}1$, gives

$$f_1(a_0, a_1, a_2, b_1, b_2) = a_1 b_1 - b_1^2 - a_0 a_2 + a_0 b_2 + 2{\cdot}1(a_1 - b_1)^2 = 0 \qquad (11\,a)$$

(2) The specification in eqn. (8 *b*), or Re $[G(j\omega_n)] = -1{\cdot}5$ at $\omega_n = 1{\cdot}9$, yields

$$f_2(a_0, a_1, a_2, b_1, b_2) = (a_2 - b_2)(a_0 - 3{\cdot}61 b_2) - b_1(a_1 - b_1 - 3{\cdot}61) - 1{\cdot}5[3{\cdot}61(a_2 - b_2)^2 + (a_1 - b_1 - 3{\cdot}61)^2] = 0 \qquad (11\,b)$$

(3) The condition in (8 *c*), or $\angle G(j\omega_n) = -180°$ at $\omega_n = 1{\cdot}9$, gives

$$f_3(a_0, a_1, a_2, b_1, b_2) = 3{\cdot}61 b_1(a_2 - b_2) + (a_0 - 3{\cdot}61 b_2)(a_1 - b_1 - 3{\cdot}61) = 0 \qquad (11\,c)$$

(4) The specification in (8 *d*), or $\phi_m = 5{\cdot}7°$ at $\omega_c = 3{\cdot}2$, yields

$$f_4(a_0, a_1, a_2, b_1, b_2) = 10{\cdot}24 b_1(a_2 - b_2) + (a_0 - 10{\cdot}24 b_2)(a_1 - b_1 - 10{\cdot}24) - 0{\cdot}319\,402\,24[(a_2 - b_2)(a_0 - 10{\cdot}24 b_2) - b_1(a_1 - b_1 - 10{\cdot}24)] = 0 \qquad (11\,d)$$

(5) The assignment in (8 *e*), or $|G(j\omega_c)| = 1$ at $\omega_c = 3{\cdot}2$, gives

$$f_5(a_0, a_1, a_2, b_1, b_2) = (a_0 - 10{\cdot}24 b_2)^2 + 10{\cdot}24 b_1^2 - 104{\cdot}8576(a_2 - b_2)^2 - 10{\cdot}24(a_1 - b_1 - 10{\cdot}24)^2 = 0 \qquad (11\,e)$$

Equation (11) is a set of high-order simultaneous non-linear algebraic equations which are very difficult to solve. Considering the availability of the computer program package (IBM 1977) (called the Z systems) in many digital computers for the solution of non-linear equations, the Newton–Raphson multidimensional method is suggested for solving these equations. However, it is well known that the Newton–Raphson method will only converge for a small range of starting values or the initial guesses. A set of good initial guesses must be determined for rapid convergence of the numerical method. Four methods are proposed for these good initial guesses.

3. The initial guess

It is well known that high-order non-linear equations have many solutions. The solution and the speed of convergence of a numerical method depend heavily on the initial guesses or the starting values. In this paper, the Newton–Raphson method is suggested for solving the non-linear equations. The following methods, depending on the applications of interest, are proposed for good initial guesses.

3.1. *Initial guess by a synthesis method*

Suppose only the dominant frequency-response data in (8) are available and an approximate transfer function $T_d^*(s)$ of the desired $T_d(s)$ in (9 *b*) is required. The $T_d^*(s)$ is

$$T_d^*(s) = \frac{a_0^* + b_1^* s + b_2^* s^2}{a_0^* + a_1^* s + a_2^* s^2 + s^3} \qquad (12)$$

where a_i^* and b_i^* are the starting values of the numerical method. The steps to obtain (12) are summarized as follows :

Step 1. Determine a second-order approximate transfer function $T_2^*(s)$ using $\phi_m = 5{\cdot}7°$ and $\omega_c = 3{\cdot}2$ rad/s in (8 *d*) and (8 *e*). This $T_2^*(s)$ is

$$T_2^*(s) = \frac{\omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2} \tag{13 a}$$

where ξ = the damping ratio and ω_n = the natural angular frequency. Two non-linear equations, which are constructed from the basic definitions of ω_c and ϕ_m, can be obtained. These non-linear equations can be converted into a single variable (ξ or ω_n) high-order equation from which the roots can be determined. Using this approach, we have $\xi = 0{\cdot}0498$ and $\omega_n = 3{\cdot}2079$. The poles that can be considered as the dominant poles of a system can be determined from the characteristic equation in (13 *a*). The dominant poles are

$$s_{1,2} = -\xi\omega_n \pm j\omega_n\sqrt{(1-\xi^2)} = -0{\cdot}1598 \pm j3{\cdot}2039 \tag{13 b}$$

Thus, (13 *a*) becomes

$$T_2^*(s) = \frac{10{\cdot}2909}{s^2 + 0{\cdot}3194s + 10{\cdot}2909} \tag{13 c}$$

Step 2. Construct a third-order approximate transfer function $T_3^*(s)$ by inserting in it a pole ($s = -p$) and modifying the term in the numerator of $T_2^*(s)$ so that the final value of the $T_3^*(s)$ equals to unity, or

$$T_3^*(s) = \frac{P\omega_n^2}{(s^2 + 2\xi\omega_n s + \omega_n^2)(s+P)} = \frac{10{\cdot}2909P}{(s^2 + 0{\cdot}3194s + 10{\cdot}2909)(s+P)} \tag{13 d}$$

The unknown constant P can be easily determined by using the condition in (8 *b*), or $\mathrm{Re}\,[G(j\omega_\pi)] = -1{\cdot}5$ where $\omega_\pi = 1{\cdot}9$. Thus, we have

$$P = 4{\cdot}5401 \tag{13 e}$$

Step 3. Establish another third-order approximate function $T_3^{**}(s)$ by inserting a zero in (13 *d*) with an unknown constant b_1^*.

$$T_3^{**}(s) = \frac{b_1^* s + P\omega_n^2}{(s^2 + 2\xi\omega_n s + \omega_n^2)(s+P)} = \frac{b_1^* s + 46{\cdot}7216}{(s^2 + 0{\cdot}3194s + 10{\cdot}2909)(s + 4{\cdot}5401)} \tag{13 f}$$

The b_1^* can be determined by using the condition in (2 *c*) and (8 *a*), or $\mathrm{Re}\,[G(j0)] = -2{\cdot}1$. The b_1^* is

$$b_1^* = 32{\cdot}4038 \tag{13 g}$$

Hence, we have

$$T_3^{**}(s) = \frac{46{\cdot}7216 + 32{\cdot}4038s}{46{\cdot}7216 + 11{\cdot}7410s + 4{\cdot}8595s^2 + s^3} \tag{13 h}$$

Equation (13 *h*) can be considered as an approximate function of (12) by assuming $b_2^* = 0$. The initial guesses in (12) are $a_0^* = 46{\cdot}7216$, $a_1^* = 11{\cdot}7410$, $a_2^* = 4{\cdot}8595$, $b_1^* = 32{\cdot}4038$, and $b_2^* = 0$. Using these constants as starting

values for the numerical method yields the desired coefficients in (9 *b*), or $a_0 = 6{\cdot}378\ 070$, $a_1 = 10{\cdot}462\ 220$, $a_2 = 1{\cdot}259\ 008$, $b_1 = 20{\cdot}556\ 61$, and $b_2 = 0{\cdot}243\ 466$. The desired transfer function is

$$T_3(s) = \frac{6{\cdot}378\ 070 + 20{\cdot}556\ 61s + 0{\cdot}243\ 466s^2}{6{\cdot}378\ 070 + 10{\cdot}462\ 220s + 1{\cdot}259\ 008s^2 + s^3} \tag{14}$$

The Newton–Raphson method (IBM 1977) converges at the 9th iteration with the error tolerance of 10^{-6}. Equation (14) has the exact frequency-response data specified in (8).

3.2. *Initial guess by complex-curve fitting and continued fraction methods*

The problem of finding unknown coefficients of a transfer function as a ratio of two frequency-dependent polynomials has been investigated by Levy (1959). His method minimizes the sum of squares of the errors at arbitrary experimental points. We present a simple method to determine the approximate coefficients of a transfer function using the real parts and imaginary parts of available limited frequency-response data. A low-order model is often determined because of data limitation. The low-order model is then expanded into a continued fraction of the Cauer second form to obtain a set of dominant quotients. Then some non-dominant quotients are inserted into the continued fraction to obtain an amplified-order model (Huang and Shieh 1976) which is the desired approximate transfer function for the use of the initial guess.

Consider the transfer function

$$T^*(s) = \frac{b_0 + b_1 s + b_2 s^2 + \ \dots\ + b_m s^m}{1 + a_1 s + a_2 s^2 + \ \dots\ + a_n s^n} \tag{15 a}$$

where a_i and b_i are unknown coefficients to be determined. Substituting $s = j\omega_k$ into (15 *a*) we have

$$\begin{aligned} T^*(j\omega_k) &= \frac{(b_0 - b_2\omega_k^2 + b_4\omega_k^4 - b_6\omega_k^6 + \ \dots) + j(b_1\omega_k - b_3\omega_k^3 + b_5\omega_k^5 - b_7\omega_k^7 + \ \dots)}{(1 - a_2\omega_k^2 + a_4\omega_k^4 - a_6\omega_k^6 + \ \dots) + j(a_1\omega_k - a_3\omega_k^3 + a_5\omega_k^5 - a_7\omega_k^7 + \ \dots)} \\ &= R(\omega_k) + jI(\omega_k) = R_k + jI_k \end{aligned} \tag{15 b}$$

when R_k and I_k are the given real and imaginary parts of the $T^*(s)$ at the available frequencies ω_k. Multiplying both sides of (15 *b*) by the common denominator and separating the real and imaginary parts, and also equating the respective real and imaginary parts, yields

$$b_0 - b_2\omega_k^2 + b_4\omega_k^4 - b_6\omega_k^6 + \ \dots\ + a_1 I_k\omega_k + a_2 R_k\omega_k^2 - a_3 I_k\omega_k^3 - a_4 R_k\omega_k^4 + \ \dots = R_k \tag{15 c}$$

and

$$b_1\omega_k - b_3\omega_k^3 + b_5\omega_k^5 - b_7\omega_k^7 + \ \dots\ - a_1 R_k\omega_k + a_2 I_k\omega_k^2 + a_3 R_k\omega_k^3 - a_4 I_k\omega_k^4 + \ \dots = I_k \tag{15 d}$$

In matrix form, (15 *c*) becomes

$$\begin{bmatrix} 1 & -\omega_1^2 & \omega_1^4 & -\omega_1^6 & . & I_1\omega_1 & R_1\omega_1^2 & -I_1\omega_1^3 & -R_1\omega_1^4 & . \\ 1 & -\omega_2^2 & \omega_2^4 & -\omega_2^6 & . & I_2\omega_2 & R_2\omega_2^2 & -I_2\omega_2^3 & -R_2\omega_2^4 & . \\ 1 & -\omega_3^2 & \omega_3^4 & -\omega_3^6 & . & I_3\omega_3 & R_3\omega_3^2 & -I_3\omega_3^3 & -R_3\omega_3^4 & . \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ 1 & -\omega_x^2 & \omega_x^4 & -\omega_x^6 & . & I_x\omega_x & R_x\omega_x^2 & -I_x\omega_x^3 & -R_x\omega_x^4 & . \end{bmatrix} \begin{bmatrix} b_0 \\ b_2 \\ b_4 \\ . \\ a_1 \\ a_2 \\ . \\ a_n \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ . \\ . \\ . \\ . \\ R_x \end{bmatrix} \tag{15 e}$$

where $x = n + m/2 + 1$ if m is even and $x = n + (m+1)/2$ if m is odd.

Substituting a_i obtained in (15 *e*) into (15 *d*), we have another matrix equation to solve for b_i, $i = 1, 3, 5, \ldots$.

$$\begin{bmatrix} \omega_1 & -\omega_1^3 & \omega_1^5 & -\omega_1^7 & . & . & . \\ \omega_2 & -\omega_2^3 & \omega_2^5 & -\omega_2^7 & . & . & . \\ \omega_3 & -\omega_3^3 & \omega_3^5 & -\omega_3^7 & . & . & . \\ . & . & . & . & . & . & . \\ \omega_y & -\omega_y^3 & \omega_y^5 & -\omega_y^7 & . & . & . \end{bmatrix} \begin{bmatrix} b_1 \\ b_3 \\ b_5 \\ . \\ b_k \end{bmatrix}$$

$$= \begin{bmatrix} ((a_0 I_1 \omega_1^0 + a_1 R_1 \omega_1^1) - (a_2 I_1 \omega_1^2 + a_3 R_1 \omega_1^3) + \ldots) \\ ((a_0 I_2 \omega_2^0 + a_1 R_2 \omega_2^1) - (a_2 I_2 \omega_2^2 + a_3 R_2 \omega_2^3) + \ldots) \\ \\ ((a_0 I_y \omega_y^0 + a_1 R_y \omega_y^1) - (a_2 I_y \omega_y^2 + a_3 R_y \omega_y^3) + \ldots) \end{bmatrix} \tag{15 f}$$

where $\omega_k^0 = 1$, $a_0 = 1$; $k = m$ and $y = (m+1)/2$ if $m =$ odd ; $k = m-1$ and $y = m/2$ if $m =$ even. In this example, the available data are

$$\left.\begin{aligned} &\omega_1 = \omega_0 = 0, \qquad R_1 = T(j0) = 1, \quad I_1 = 0 \\ &\omega_2 = \omega_\pi = 1{\cdot}9, \quad R_2 = \mathrm{Re}\left[\frac{G(j\omega_\pi)}{1+G(j\omega_\pi)}\right] = 2{\cdot}9684, \\ &\qquad\qquad I_2 = I_\mathrm{m}\left[\frac{G(j\omega_\pi)}{1+G(j\omega_\pi)}\right] = -0{\cdot}0252 \\ &\omega_3 = \omega_\mathrm{c} = 3{\cdot}2, \quad R_3 = \mathrm{Re}\left[\frac{G(j\omega_\mathrm{c})}{1+G(j\omega_\mathrm{c})}\right] = 0{\cdot}{\,}351, \\ &\qquad\qquad I_3 = I_\mathrm{m}\left[\frac{G(j\omega_\mathrm{c})}{1+G(j\omega_\mathrm{c})}\right] = -10{\cdot}4316 \end{aligned}\right\} \tag{16}$$

Since only three values are available, the approximate function $T_2^*(s)$ is

$$T_2^*(s) = \frac{b_0 + b_1 s}{1 + a_1 s + a_2 s^2} \tag{17 a}$$

Substituting the data at ω_1, and ω_2, and ω_3 in (16) into (15 *e*) yields $b_0 = 1$, $a_1 = 0{\cdot}0388$, and $a_2 = 0{\cdot}1839$. Then substituting a_i and the data at ω_3 into

(15 *f*) gives $b_1 = 2{\cdot}8907$. Because the desired approximate function in (12) is a third-order function, $T_2^*(s)$ should be amplified by using the continued fraction method (Huang and Shieh 1976) as follows.

$T_2^*(s)$ is first expanded into a continued fraction of the Cauer second form to obtain a set of dominant quotients : $h_1 = 1$, $h_2 = -0{\cdot}3507$, $h_3 = -0{\cdot}9651$, and $h_4 = 16{\cdot}0725$. Then the order of $T_2^*(s)$ is amplified to the third order by inserting non-dominant quotients $h_5 = 100$ and $h_6 = 0{\cdot}1$, or

$$T_2^*(s) = \frac{1 + 2{\cdot}8907s}{1 + 0{\cdot}0388s + 0{\cdot}1839s^2} = \cfrac{1}{h_1 + \cfrac{s}{h_2 + \cfrac{s}{h_3 + \cfrac{s}{h_4}}}} \simeq \cfrac{1}{h_1 + \cfrac{s}{h_2 + \cfrac{s}{h_3 + \cfrac{s}{h_4 + \cfrac{s}{h_5 + \cfrac{s}{h_6}}}}}}$$

$$= T_3^*(s) = \frac{54{\cdot}3885 + 162{\cdot}6914s + 15{\cdot}8219s^2}{54{\cdot}3885 + 7{\cdot}5839s + 10{\cdot}2146s^2 + s^3} \qquad (17\,b)$$

Huang and Shieh (1976) have shown that the amplified-order model is a good approximation of the original low-order model if the inserted positive quotients $h_i \gg 1$ and $h_{i+1} \ll 1$ where i is an odd number. Using the coefficients in (17 *b*) as initial guesses we have the desired coefficients in (14) at the 15th iteration (IBM 1977) with the error tolerance of 10^{-6}.

If much experimental frequency-response data, including the dominant data of a system, is available and the transfer function of the original system is required, this is an identification problem. In this case, a set of non-linear equations, based on the basic definitions of the dominant data, can be constructed and can be solved by the Newton–Raphson method. The initial guess can be determined by using the dominant data and others in (15). Since many data are available, a high-order approximate transfer function can be determined. Therefore, the use of the continued fraction method (Huang and Shieh 1976) is not necessary.

When a high-order transfer function of a system is given and various reduced-order transfer functions are required, this is a model reduction problem. In the frequency domain, numerous methods (Chen and Shieh 1969, Shieh and Goldman 1974, Hutton and Friedland 1975, Sharnash 1975, Lal and Van Valkenburg 1976) have been proposed for model reduction. The continued fraction methods (Chen and Shieh 1969, Shieh and Goldman 1974), the Routh approximation method (Hutton and Friedland 1975), the time-moment matching method (Shamash 1975), and the frequency-moment matching method (Lal and Van Valkenburg 1976) are the typical examples. These methods have been critically compared by Decoster and Cauwenberghe (1976). The new method presented in this paper can be used to obtain the reduced-order models which have the exact dominant frequency-response data as those of the original one. This method can be called a dominant frequency-response data matching method. The procedure is as follows.

Step 1. Plot the frequency-response curves to determine the data at the dominant frequencies ω_0, ω_π, ω_c, ω_{c1}, ω_{c2}, ω_p, and ω_h.

Step 2. Formulate a low-order model with unknown coefficients, and write a set of non-linear equations based on the basic definitions of the data at dominant frequencies.

Step 3. Determine a set of good starting values by using the synthesis method or the complex curve fitting method, and solve the non-linear equation by using the Newton–Raphson method. Thus, reduced-order models can be determined. Comparing the reduced-order models obtained from the proposed method with those of the existing methods (Chen and Shieh 1969, Shieh and Goldman 1974, Hutton and Friedman 1975, Shamash 1975, Lal and Van Valkenburg 1976), we observe that the model obtained in this paper is superior to existing methods in that the reduced model has the exact dominant frequency response as the original. As a result, an engineer can design a control system more efficiently in the frequency domain.

Since the original high-order transfer function is available, an existing method (Chen and Shieh 1969) can be applied and modified to obtain an approximate transfer function for the determination of the initial guess. Two additional methods for initial guess determination are as follows.

(3) Initial guess by a continued fraction method (Chen and Shieh 1969).

Consider the high-order transfer function in (1 *a*). The function can be expanded into a continued fraction and various reduced models obtained by discarding some of the quotients, or

$$T(s)=\frac{b_0+b_1s+\ \dots\ +b_ms^m}{a_0+a_1s+\ \dots\ +a_ns^n}=\frac{n(s)}{d(s)} \tag{18 a}$$

$$=\cfrac{1}{h_1+\cfrac{s}{h_2+\cfrac{s}{\ddots}}} \tag{18 b}$$

$$\simeq\frac{1}{h_1+\dfrac{s}{h_2}}=\frac{h_2}{h_1h_2+s} \tag{18 c}$$

$$\simeq\cfrac{1}{h_1+\cfrac{s}{h_2+\cfrac{s}{h_3+\dfrac{s}{h_4}}}}=\frac{h_2h_3h_4+(h_2+h_4)s}{h_1h_2h_3h_4+(h_1h_2+h_1h_4+h_3h_4)s+s^2} \tag{18 d}$$

$$\simeq\ \dots$$

Using the coefficients of the approximate model in (18) as the initial guess for the numerical method, we have the desired reduced model. However, the

approximate model in (18) may be unstable even if the original system is stable. The continued fraction method (Chen and Shieh 1969) can be modified by the following new method.

(4) Initial guess by a mixed method of the continued fraction approach and Gustafson's (1965) method.

Assume the reduced model of the original system in (18 *a*) is

$$T_p^*(s)=\frac{b_0^*+b_1^* s+\ \dots\ +b_{p-1}^* s^{p-1}}{a_0^*+a_1^* s+\ \dots\ +a_p^* s^p}=\frac{n^*(s)}{d^*(s)},\quad a_p=1 \tag{19 a}$$

A matrix equation (Chen and Shieh 1970) can be constructed from the dominant quotients h_i, $i=1, 2, \dots, p$, obtained in (18 *b*) and the unknown coefficients a_i^* and b_i^* in (19 *a*) as

$$[b]=[H][a] \tag{19 b}$$

where

$$[a]^{\mathrm{T}}=[a_0^*, a_1^*, \dots, a_{p-1}^*] \tag{19 c}$$

$$[b]^{\mathrm{T}}=[b_0^*, b_1^*, \dots, b_{p-1}^*] \tag{19 d}$$

$$[H]=[H_2]^{-1}[H_1] \tag{19 e}$$

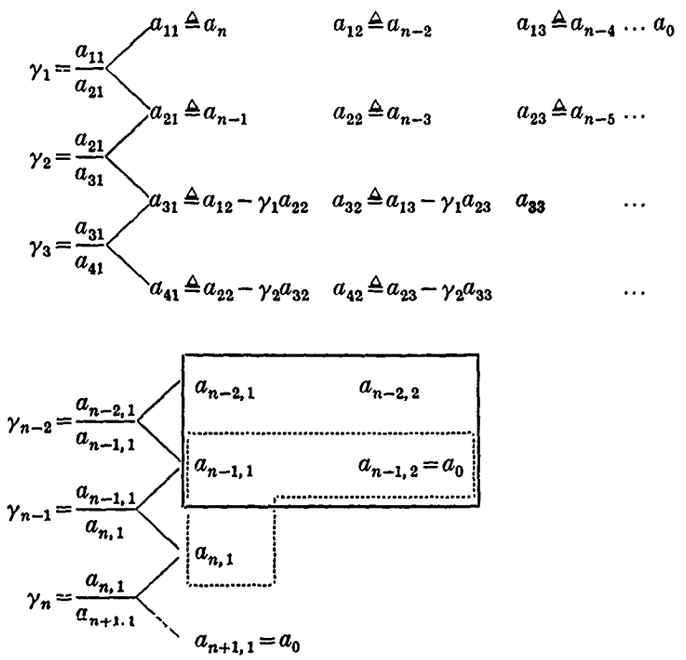
where T designates transpose,

$$[H_2]=\begin{bmatrix} h_1 & 0 & 0 & . & 0 & 0 \\ 1 & h_2 & 0 & . & 0 & 0 \\ 0 & 1 & h_3 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1 & h_p \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & . & 0 & 0 \\ 0 & h_1 & 0 & . & 0 & 0 \\ 0 & 1 & h_2 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1 & h_{p-1} \end{bmatrix}\cdots\begin{bmatrix} 1 & 0 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 0 & h_1 \end{bmatrix}$$

$$[H_1]=\begin{bmatrix} 1 & 0 & 0 & . & 0 & 0 \\ 0 & h_2 & 0 & . & 0 & 0 \\ 0 & 1 & h_3 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1 & h_p \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 \\ 0 & 0 & h_2 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 1 & h_{p-1} \end{bmatrix}\cdots\begin{bmatrix} 1 & 0 & 0 & . & 0 & 0 \\ 0 & 1 & 0 & . & 0 & 0 \\ 0 & 0 & 1 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & 0 & h_2 \end{bmatrix}$$

The a_i^* in (19 *c*) can be determined from the coefficients of the polynomial that is obtained from the product of the dominant eigenvalues of the $d(s)$ in (18 *a*). When the dominant poles of $d(s)$ cannot be clearly identified or the poles of $d(s)$ are not available, the paper and pencil method suggested by Gustafson (1965) can be applied to construct the $d^*(s)$ or to determine a_i^* in (19 *c*). Then, substituting the a_i^* into (19 *b*) yields the required $n^*(s)$ or b_i^* in (19 *a*). The steps determine the $d^*(s)$ are shown as follows.

Step 1. Construct a Routh (1877) array using the coefficients a_i of $d(s)$ and the Routh algorithm. The a_i are expressed by double-subscripted notation $a_{i,j}$ for obtaining the general algorithm. The Routh array is



(20 *a*)

In general $a_{i,j} = a_{i-2,j+1} - \gamma_{i-2} a_{i-1,j+1}$; $i = 1, 2, \ldots, j = 3, 4, \ldots$

$$\gamma_i = a_{i,1}/a_{i+1,1} \tag{20 b}$$

Step 2. Construct various approximate low-order polynomials $d_i^*(s)$ from the last row and the next to last row, and so on in the Routh array.

For example, the ith order approximate equations are

$$d_1^*(s) = a_{n,1} s + a_{n+1,1} = a_{n,1} s + a_0 = 0 \quad \text{when } i = 1 \tag{20 c}$$

$$d_2^*(s) = a_{n-1,1} s^2 + a_{n,1} s + a_{n-1,2} = a_{n-1,1} s^2 + a_{n,1} s + a_0 = 0 \quad \text{when } i = 2 \tag{20 d}$$

and

$$\begin{aligned} d_3^*(s) &= a_{n-2,1} s^3 + a_{n-1,1} s^2 + a_{n-2,2} s + a_{n-1,2} \\ &= a_{n-2,1} s^3 + a_{n-1,1} s^2 + a_{n-2,2} s + a_0 = 0 \quad \text{when } i = 3 \end{aligned} \tag{20 e}$$

Since the original system is asymptotically stable, all γ_i are positive values. The approximate polynomials $d_i^*(s)$ are always the Hurwitz polynomials. Moreover, Gustafson (1965) has shown that relationships exist between the coefficients of $d_i^*(s)$ and the time-domain moments. The normalized polynomials can be determined by dividing each coefficient in $d_i^*(s)$ by the coefficient of the highest order term in s. The approximate transfer function $T_p^*(s)$

in (19 *a*) can be considered as a reduced-order model of the original high-order system. In this paper, we use it as the initial guess for the numerical method for determining the reduced order model that has the exact dominant frequency-response data as the original system.

4. An illustrative example

Consider the unit ratio feedback closed-loop transfer function of a stabilized real missile system (Bosley 1977)

$$T(s)=\frac{k'(b'_0+b'_1 s+ \ \dots \ +b'_5 s^5)}{a_0+a_1 s+ \ \dots \ +a_{11}s^{11}} \qquad (21\ a)$$

where

$$\begin{aligned}
a_0&=8{\cdot}802\ 158\ 509\times 10^{18}, & a_1&=2{\cdot}419\ 047\ 424\times 10^{19}\\
a_2&=2{\cdot}911\ 920\ 56\times 10^{18}, & a_3&=2{\cdot}420\ 405\ 431\times 10^{18}\\
a_4&=6{\cdot}667\ 397\ 031\times 10^{16}, & a_5&=9{\cdot}749\ 923\ 212\times 10^{14}\\
a_6&=9{\cdot}360\ 329\ 977\times 10^{12}, & a_7&=6{\cdot}231\ 675\ 318\times 10^{10}\\
a_8&=2{\cdot}976\ 950\ 696\times 10^{8}, & a_9&=9{\cdot}316\ 239\ 04\times 10^{5}\\
a_{10}&=1{\cdot}923\ 554\times 10^{3}, & a_{11}&=1
\end{aligned}$$

and

$$\begin{aligned}
k'&=1{\cdot}494\ 523\ 312\times 10^{11}\\
b'_0&=5{\cdot}889\ 609\ 375\times 10^{7}, & b'_1&=3{\cdot}084\ 598\ 703\times 10^{8}\\
b'_2&=1{\cdot}958\ 045\ 299\times 10^{7}, & b'_3&=3{\cdot}357\ 065\ 095\times 10^{5}\\
b'_4&=1{\cdot}715\ 193\ 3\times 10^{3}, & b'_5&=1
\end{aligned}$$

The second order and the third order reduced-order models which have some of the dominant frequency-response data of the original system are required. The open-loop transfer function $G(s)$ of the system is

$$G(s)=\frac{k(e_0+e_1 s+ \ \dots \ +e_5 s^5)}{s(g_0+g_1 s+ \ \dots \ +g_{10}s^{10})} \qquad (21\ b)$$

where

$$\begin{aligned}
g_0&=-2{\cdot}190\ 952\ 724\ 6\times 10^{19}, & g_1&=-1{\cdot}442\ 378\ 55\times 10^{16}\\
g_2&=2{\cdot}370\ 233\ 311\times 10^{18}, & g_3&=6{\cdot}641\ 763\ 067\times 10^{16}\\
g_4&=9{\cdot}748\ 428\ 689\times 10^{14}, & g_5&=9{\cdot}360\ 329\ 977\times 10^{12}\\
g_6&=6{\cdot}231\ 675\ 318\times 10^{10}, & g_7&=2{\cdot}976\ 950\ 696\times 10^{8}\\
g_8&=9{\cdot}316\ 239\ 04\times 10^{5}, & g_9&=1{\cdot}923\ 554\times 10^{3}\\
g_{10}&=1
\end{aligned}$$

and

$$\begin{aligned}
k&=1{\cdot}494\ 523\ 312\times 10^{11}\\
e_0&=5{\cdot}889\ 609\ 375\times 10^{7}, & e_1&=3{\cdot}084\ 598\ 703\times 10^{8}\\
e_2&=1{\cdot}958\ 045\ 299\times 10^{7}, & e_3&=3{\cdot}357\ 065\ 095\times 10^{5}\\
e_4&=1{\cdot}715\ 193\ 3\times 10^{3}, & e_5&=1
\end{aligned}$$

Note that $G(s)$ is a non-minimum phase function ; its Nyquist plot is shown in Fig. 1. The dominant frequency-response data are chosen and given in (8). The set of non-linear equations are shown in (11). The initial guesses shown in (13 *h*) and (17 *b*) yields the required third-order reduced model in (14), or

$$T_3^*(s) = \frac{0{\cdot}243\ 466s^2 + 20{\cdot}556\ 61s + 6{\cdot}378\ 07}{s^3 + 1{\cdot}259\ 008s^2 + 10{\cdot}462\ 22s + 6{\cdot}378\ 07} \tag{22 a}$$

If the continued fraction method (Chen and Shieh 1969) in (18) is used, the approximate reduced model is

$$T_{3c}^*(s) = \frac{0{\cdot}6920s^2 + 19{\cdot}4692s + 3{\cdot}7376}{s^3 + 0{\cdot}9488s^2 + 10{\cdot}1661s + 3{\cdot}7376} \tag{22 b}$$

Using the coefficients in (22 *b*) as starting values for solving the non-linear equations in (11) yields the desired coefficients in (22 *a*) at the eighth iteration (IBM 1977) with the error tolerance of 10^- . If the mixed method in (19) and (20) is used, the normalized approximate denominator in (20 *e*) is

$$d_3^*(s) = s^3 + 0{\cdot}9524s^2 + 10{\cdot}1924s + 3{\cdot}7455 \tag{22 c}$$

The $n_3^*(s)$ obtained from (19) is

$$n_3^*(s) = 0{\cdot}7066s^2 + 19{\cdot}5155s + 3{\cdot}7455 \tag{22 d}$$

The approximate transfer function by the mixed method is

$$T_{3m}^*(s) = \frac{0{\cdot}7066s^2 + 19{\cdot}5155s + 3{\cdot}7455}{s^3 + 0{\cdot}9524s^2 + 10{\cdot}1924s + 3{\cdot}7455} \tag{22 e}$$

If the coefficients in (22 *e*) are used as starting values, the Newton–Raphson method (IBM 1977) will converge to the desired solution in (22 *a*) at the eighth iteration with the error tolerance of 10^{-6}. The unit step response curves of various reduced models and the original system are compared in Fig. 2. All three reduced-order models give very satisfactory approximate time response curves. However, only the $T_3^*(s)$ in (22 *a*), which uses the method of dominant frequency-response data matching, has the exact dominant frequency-response data as the original system.

If $\omega_c = 3{\cdot}2$ rad/s, $\phi_m = 5{\cdot}7°$ and $\mathrm{Re}\,[G(j0)] = -2{\cdot}1$ are chosen as the dominant data, the second-order reduced model obtained by the proposed method is

$$T_2^*(s) = \frac{3{\cdot}339\ 517s + 9{\cdot}224\ 24}{s^2 + 0{\cdot}302\ 806s + 9{\cdot}224\ 24} \tag{23 a}$$

The approximate reduced models by the continued fraction method and the mixed method are :

$$T_{2c}^*(s) = \frac{24{\cdot}7981s + 4{\cdot}8122}{s^2 + 12{\cdot}8201s + 4{\cdot}8122} \tag{23 b}$$

and

$$T_{2m}^*(s) = \frac{16{\cdot}3618s + 3{\cdot}9328}{s^2 + 6{\cdot}5726s + 3{\cdot}9328} \tag{23 c}$$

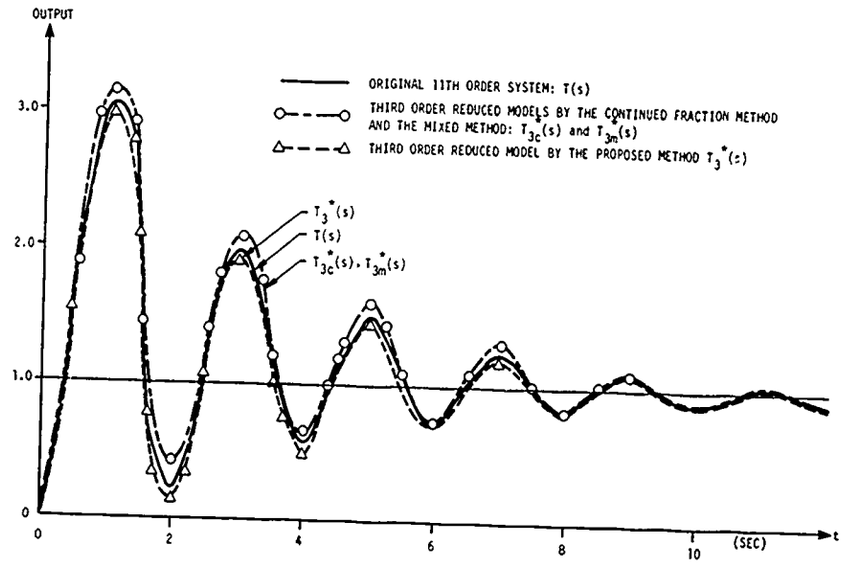


Figure 2. Time responses or original and third-order reduced models.

The unit-step time response curves of various reduced-order models $T_3^*(s)$, $T_2^*(s)$, $T_{2c}^*(s)$, and $T_{2m}^*(s)$ are compared in Fig. 3. It is observed that $T_2^*(s)$ gives better approximation in the transient response than $T_{2c}^*(s)$ and $T_{2m}^*(s)$.

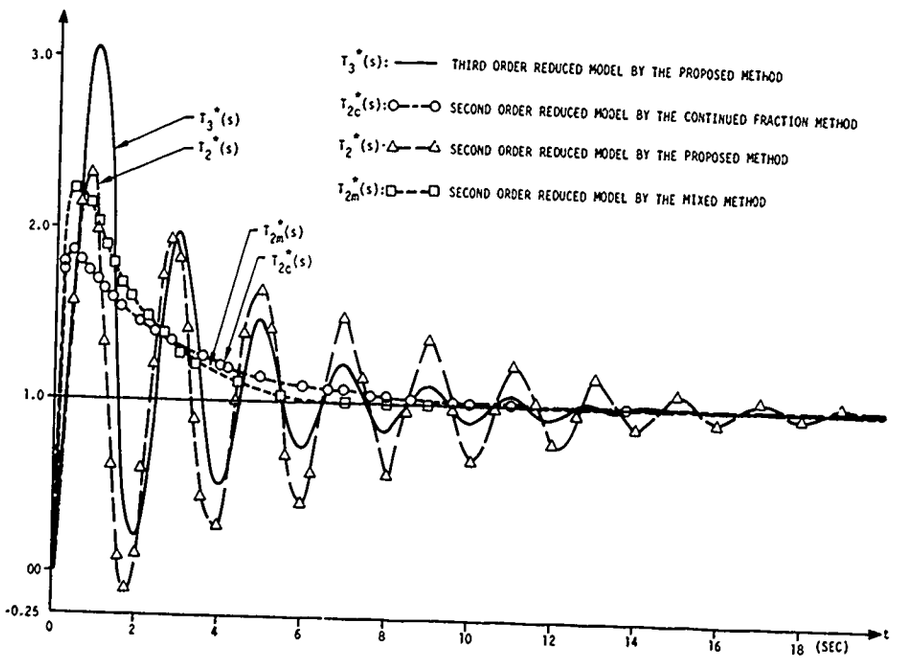


Figure 3. Time responses of third- and second-order reduced models.

5. Conclusion

A basic method has been developed for modelling transfer function using dominant frequency-response data. When the specifications of the design goals of a control system are assigned, the proposed method gives the standard transfer function. Thus, the design processes in the frequency domain can be significantly simplified. When the experimental frequency-response data of a system are available, the proposed method can be used to identify the transfer function of the original system. Also, if a high-order transfer function is given, various low-order models can be determined. The reduced models have the same dominant characteristics of the original system. Four methods have been proposed for estimating the good starting values for the solution of non-linear equations. The new dominant frequency-response data matching method, and the new mixed method that has the advantages of both continued fraction method of Chen and Shieh (1969) and the paper and pencil method of Gustafson (1965) have been developed for model reduction.

Acknowledgments

This work was supported in part by U.S. Army Missile Research and Development Command, DAAK 40-79-C-0061 and U.S. Army Research Office DAAG 29-77-G-0143.

References

Axelby, G. S., 1960, *Proc. 1st I.F.A.C.*, pp. 68–74.

Bode, H. W., 1954, *Network Analysis and Feedback Amplifier Design* (New York: Van Nostrand).

Bosley, J. T., 1977, U.S. Army Missile Command, DAAK40-77-C-0048, TGT-001.

Carnahan, B., Luther, H. A., and Wilkes, J. O., 1969, *Applied Numerical Methods* (New York: Wiley), pp. 319–329.

Chen, K., 1957, *A.I.E.E. Trans. Applic. Ind.* 76, 80.

Chen, C. F., and Shieh, L. S., 1968, *Int. J. Control*, **8**, 561; 1970, *Ibid.*, **11**, 717.

Decoster, M., and Van Cauwenberghe, A. R., 1976, *Journal* A, **17**, No. 2, 68; No. 3, 125.

Del Toro, V., and Parker, S., 1960, *Principles of Control Systems Engineering* (New York: McGraw-Hill), pp. 665–669, 278–302.

Evans, W. R., 1953, *Control System Dynamics* (New York: McGraw-Hill).

Gibson, J. E., and Rekasius, Z. V., 1961, *A.I.E.E. Trans. Applic. Ind.*, p. 65 (May, part II).

Gustafson, R. D., 1965, *Proc. JACC*, p. 301.

Huang, C. J., and Shieh, L. S., 1976, *Int. J. Systems Sci.*, **7**, 241.

Hutton, M. F., and Friedland, B., 1975, *I.E.E.E. Trans. autom. Control*, **20**, 329.

IBM, 1977, IBM S/370-360 Reference Manual IMSL (The International Mathematical and Statistical Library).

Lal, M., and Van Valkenburg, M. E., 1976, *9th Annual Asilomar Conference on Circuits, Systems, and Computers*, pp. 242–246.

Levy, E. C., 1959, *I.R.E. Trans. Autom. Control*, **4**, 37.

Nyquist, H., 1932, *Bell Syst. tech. J.*, **2**, 126.

Routh, E. J., 1877, *A Treatise on the Stability of a Given State of Motion* (London: Macmillan).

Seshadri, V., Rao, V. R., Eswaran, C., and Eappen, S., 1969, *Proc. Inst. elect. electron. Engrs*, **57**, 1321.

Shamash, Y., 1975, *Int. J. Control*, **21**, 257.

SHIEH, L. S., and GOLDMAN, M. J.. 1974, *I.E.E.E. Trans. Syst. Man Cybernetics*, **4**, 584.

THALER, G. J., 1973, *Design of Feedback Systems* (Pa : Dowden, Hutchinson & Ross).

TRUXAL, J. G., 1955, *Control System Synthesis* (New York : McGraw-Hill), pp. 76–87.

WAKELAND, W. R., 1976, *I.E.E.E. Trans. autom. Control*, **21**, 771.

Analysis and synthesis of matrix transfer functions using the new block-state equations in block-tridiagonal forms

L.S. Shieh, Ph.D., and A. Tajvari

Indexing terms: *Linear network analysis, Linear systems, Matrix algebra, Polynomials, Stability criteria, Transfer functions*

Abstract: A new block-Routh array with block-Routh algorithm is developed to extract the greatest common matrix polynomial of two matrix polynomials that are not coprime, and to construct a block-transformation matrix that transforms a block-state equation from a block-companion form to a block-tridiagonal form. The newly developed block-state equation in the block-tridiagonal form is a minimal realisation of a matrix-transfer function. Also, the block-state equation is used to synthesise a driving-point impedance matrix. A stability criterion is then derived to test the stability of a class of matrix transfer functions.

1 Introduction

The accurate description of linear, time-invariant circuits and systems in the time domain may result in m nth-degree coupled differential equations, or an nth-degree matrix differential equation with $m \times m$ matrix coefficients, as

$$\sum_{i=1}^{n+1} A_i D^{i-1} x_0(t) = r(t) \tag{1a}$$

$$y(t) = \sum_{i=1}^{n} B_i D^{i-1} x_0(t) \tag{1b}$$

and

$$D^{i-1} x_0(0) = [\alpha_i] \qquad i = 1, 2, \ldots, n \tag{1c}$$

where ~~$Y(s)$ and $R(s)$ are the Laplace transforms of $y(t)$ and vector, and~~ $x_0(t)$ is an $m \times 1$ state vector. A_i and B_i are $m \times m$ matrix coefficients and the differential operator $D = d/dt$. When each initial vector $[\alpha_i]$ is an $m \times 1$ null vector, $0_{m \times 1}$, the corresponding frequency domain representation of the same system is an nth-degree matrix transfer function written as

$$Y(s) = T(s)R(s) \tag{2a}$$

where $Y(s)$ and $R(s)$ are the Laplace transforms of $y(t)$ and $r(t)$, respectively. The matrix transfer function $T(s)$ is

$$T(s) = D_2(s)D_1(s)^{-1} \tag{2b}$$

where

$$D_1(s) = A_{n+1}s^n + A_n s^{n-1} + \ldots + A_2 s + A_1$$

and

$$D_2(s) = B_n s^{n-1} + B_{n-1} s^{n-2} + \ldots + B_2 s + B_1$$

When the matrix polynomials $D_1(s)$ and $D_2(s)$ are right coprime[1] and $A_{n+1} = I_m$, the corresponding first-degree state equation in the controllable block phase-variable form[2] (or in the controllable block companion form) is

$$\dot{x} = Ax + Br \tag{3a}$$

$$y = Cx \qquad x(0) = 0_{nm \times 1} \tag{3b}$$

where

$$A = \left[\begin{array}{ccc|c} 0_m & I_m & 0_m & 0_m \\ 0_m & 0_m & I_m & 0_m \\ \hline -A_1 & -A_2 & -A_3 & -A_n \end{array}\right],$$

$$B = \begin{bmatrix} 0_m \\ 0_m \\ \vdots \\ I_m \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$C = [B_1 \quad B_2 \quad B_3 \ldots B_n]$$

The block elements A_i, 0_m, I_m, B_i and $0_{nm \times 1}$ are $m \times m$ constant matrices, $m \times m$ null matrix, $m \times m$ identity matrix, $m \times m$ constant matrices and $nm \times 1$ null vector, respectively. The state vector x consists of n block vectors (x_i, $i = 1, 2, \ldots, n$). Each $m \times 1$ block vector x_i consists of m state variables. The state vector x is defined as a block vector, and its co-ordinates as block co-ordinates. As a result, the state equation in eqn. 3 is defined as a block-state equation in the phase-variable block co-ordinates. Without considering the special structure (the block-companion form) of the state equation in eqn. 3, the same vector x can be observed as a vector with nm state variables in general co-ordinates. When a dynamic system is formulated in a matrix differential equation or a matrix transfer function, it is more natural and convenient to analyse and synthesise such a system using the block state formulations in block co-ordinates than in general co-ordinates. In this paper, all derivations are based on the block state-space formulations in various block co-ordinates.

The objectives of this paper are described as follows:

(i) Construct new block-Routh array and block-Routh algorithm for extracting the greatest common matrix polynomial of two matrix polynomials that are not coprime, and for establishing a new block-state equation in a block-tridiagonal form, using a simple linear-block transformation.

(ii) Apply the obtained block-state equations for finding the minimal realisations of a matrix-transfer function, for synthesising the driving-point impedance matrix, and for determining the stability of a class of matrix-transfer functions.

Paper 507D, received 4th September and in revised form 3rd December 1979

Prof. Shieh and Mr. Tajvari are with the Department of Electrical Engineering, University of Houston, Houston, Texas 77004, USA

2 Construction of a new block-tridiagonal matrix

Jury,[3] Anderson,[4] Barnett[5] and Takahashi[6] have shown that a state equation in the phase-variable form can be transformed into a scalar-tridiagonal form with complex elements by extending the ideas of Chen and Chu,[7] Barnett and Storey,[8] Loo[9] and Power.[10] However, their methods deal only with single-input, single-output systems, and involve complex numbers. For multivariable systems, Shieh *et al.*[2,11] have developed linear block transformations that transform the block-state equations from the block-companion forms into the block-Schwarz forms, which are the special block-tridiagonal forms. However, their methods are restricted to a characteristic matrix polynomial and not to the matrix transfer function. A new block-Routh array with block-Routh algorithm is developed to construct a simple linear-block transformation that transforms a block-state equation from the block-companion form to a block-tridiagonal form. The construction of the block transformation involves the real matrix coefficients in both numerator and denominator matrix polynomials of a matrix transfer function.

In order to derive the recurrence algorithms, the block elements A_i and B_i in eqn. 2 are expressed by the double subscripted block elements $D_{1,j}$ and $D_{2,j}$, respectively, as

$$D_{1,j} = A_{n+2-j} \qquad j = 1,2,\ldots,n+1 \tag{4a}$$

$$D_{2,j} = B_{n+1-j} \qquad j = 1,2,\ldots,n \tag{4b}$$

Rewriting eqn. 2 yields

$$\begin{aligned} T(s) &= D_2(s)D_1(s)^{-1} \\ &= [B_n s^{n-1} + B_{n-1}s^{n-2} + \ldots + B_1] \\ &\quad \times [A_{n+1}s^n + A_n s^{n-1} + \ldots + A_1]^{-1} \\ &= [D_{21}s^{n-1} + D_{22}s^{n-2} + \ldots + D_{2,n}] \\ &\quad \times [D_{11}s^n + D_{12}s^{n-1} + \ldots + D_{1,n+1}]^{-1} \end{aligned} \tag{5}$$

When $D_2(s)$ and $D_1(s)$ are right coprime and $D_{11} = A_{n+1} = I_m$, the block-state equation in eqn. 3 can be transformed into a block-tridiagonal form using the following new block transformation:

$$x = T_1^{-1}z \tag{6}$$

where

$$T_1 = \left[\begin{array}{cc|cccc} D_{2n,1} & 0_m & 0_m & 0_m & 0_m & 0_m \\ D_{2n-2,2} & D_{2n-2,1} & 0_m & 0_m & 0_m & 0_m \\ \hline D_{8,n-3} & D_{8,n-4} & D_{81} & 0_m & 0_m & 0_m \\ D_{6,n-2} & D_{6,n-3} & D_{62} & D_{61} & 0_m & 0_m \\ D_{4,n-1} & D_{4,n-2} & D_{43} & D_{42} & D_{41} & 0_m \\ D_{2,n} & D_{2,n-1} & D_{24} & D_{23} & D_{22} & D_{21} \end{array}\right] \tag{7}$$

The new block-state equation in the block-tridiagonal form is

$$\dot{z} = T_1AT_1^{-1}z + T_1Br = G_1z + E_1r \tag{8a}$$

$$y = CT_1^{-1}z = F_1z \tag{8b}$$

where the system matrix $G_1 (= T_1AT_1^{-1})$ is the block-tridiagonal matrix and is written as

$$G_1 = \left[\begin{array}{cc|cccc} -(H_{2n}K_{2n-1})^{-1} & K_{2n-1}^{-1} & 0_m & 0_m & 0_m & 0_m \\ (H_{2n-2}K_{2n-3})^{-1} & -(H_{2n-2}K_{2n-3})^{-1} & 0_m & 0_m & 0_m & 0_m \\ \hline 0_m & 0_m & -(H_8K_7)^{-1} & K_7^{-1} & 0_m & 0_m \\ 0_m & 0_m & (H_6K_5)^{-1} & -(H_6K_5)^{-1} & K_5^{-1} & 0_m \\ 0_m & 0_m & 0_m & (H_4K_3)^{-1} & -(H_4K_3)^{-1} & K_3^{-1} \\ 0_m & 0_m & 0_m & 0_m & (H_2K_1)^{-1} & -(H_2K_1)^{-1} \end{array}\right] \tag{8c}$$

$$\begin{aligned} E_1' &= (T_1B)' \\ &= [0_m \quad 0_m \text{ ----- } 0_m \quad 0_m \quad 0_m \quad (K_1^{-1})'] \end{aligned} \tag{8d}$$

$$\begin{aligned} F_1 &= CT_1^{-1} \\ &= [0_m \quad 0_m \text{ ----- } 0_m \quad 0_m \quad 0_m \quad I_m] \end{aligned} \tag{8e}$$

The ' in eqn. 8*d* designates transpose. The block elements $D_{i,j}$ in eqn. 7 and K_i and H_i in eqn. 8 can be obtained from the following new block-Routh array with block-Routh algorithm which is different from the matrix-Routh array with matrix-Routh algorithm developed by Shieh and Gaudiano.[12]

The block-Routh array is

$$
\begin{array}{lllll}
 & D_{11} = A_{n+1} & D_{12} = A_n & D_{13} = A_{n-1} & D_{1,n+1} = A_1 \\
K_1 = D_{11}D_{21}^{-1} & & & & \\
 & D_{21} = B_n & D_{22} = B_{n-1} & D_{23} = B_{n-2} & \cdots \\
H_2 = D_{21}D_{31}^{-1} & & & & \\
K_3 = D_{21}D_{41}^{-1} & D_{31} \triangleq D_{12} - K_1 D_{22} & D_{32} \triangleq D_{13} - K_1 D_{23} & D_{33} & \cdots \\
 & D_{41} \triangleq D_{22} - H_2 D_{32} & D_{42} \triangleq D_{23} - H_2 D_{33} & D_{43} & \cdots \\
H_4 = D_{41}D_{51}^{-1} & & & & \\
K_5 = D_{41}D_{61}^{-1} & D_{51} \triangleq D_{32} - K_3 D_{42} & D_{52} \triangleq D_{23} - K_3 D_{43} & D_{53} & \cdots \\
 & D_{61} \triangleq D_{42} - H_4 D_{52} & D_{62} \triangleq D_{43} - H_4 D_{53} & D_{63} & \cdots \\
 & \cdots & \cdots & & \\
 & D_{2n-2,1} & D_{2n-2,2} & & \\
H_{2n-2} = D_{2n-2,1}D_{2n-1,1}^{-1} & & & & \\
K_{2n-1} & D_{2n-1,1} & D_{2n-1,2} & & \\
= D_{2n-2,1}D_{2n,1}^{-1} & D_{2n,1} \triangleq D_{2n-2,2} - H_{2n-2}D_{2n-1,2} & & & \\
H_{2n} = D_{2n,1}D_{2n+1,1}^{-1} & & & & \\
 & D_{2n+1,1} \triangleq D_{2n-2,2} & & &
\end{array}
\qquad (9a)
$$

The general block-Routh algorithm is

(i)

$$
\begin{aligned}
K_1 &= D_{11}D_{21}^{-1} \quad \text{rank } D_{21} = m \\
D_{3,j} &= D_{1,j+1} - K_1 D_{2,j+1} \quad j = 1, 2, \ldots, n \\
H_2 &= D_{21}D_{31}^{-1} \quad \text{rank } D_{31} = m \\
D_{4,j} &= D_{2,j+1} - H_2 D_{3,j+1} \quad j = 1, 2, \ldots, n-1
\end{aligned}
\qquad (9b)
$$

(ii)

$$
\left.
\begin{aligned}
K_{i+1} &= D_{i,1}D_{i+2,1}^{-1} \\
&\quad \text{rank } D_{i+2,1} = m \\
D_{i+3,j} &= D_{i,j+1} \\
&\quad - K_{i+1}D_{i+2,j+1} \\
H_{i+2} &= D_{i+2,1}D_{i+3,1}^{-1} \\
&\quad \text{rank } D_{i+3,1} = m \\
D_{i+4,j} &= D_{i+2,j+1} \\
&\quad - H_{i+2}D_{i+3,j+1} \\
D_{2n+2,j} &= 0_m
\end{aligned}
\right\}
\begin{aligned}
&j = 1, 2, 3, \ldots \\
&i = 2, 4, 6, \ldots, 2n-2
\end{aligned}
\qquad (9c)
$$

The construction of the block-Routh array in eqn. 9 can be described as follows: arrange the matrix coefficients of the given matrix polynomial $D_1(s)$ in eqn. 2 in the first row of eqn. 9 and the $D_2(s)$ in the second row. A new matrix K_1 is obtained by the matrix multiplication $D_{11}D_{21}^{-1}$ where D_{11} and D_{21} are the block elements in the first column of the array.

The block elements in the third row are generated from the K_1 and the block elements in the first two rows as follows: first, each block element in the second row is premultiplied by K_1. Then, subtract each resulting block from each block element in the first row. Finally, shift each block element so obtained one column left and drop the zero-first block element to form the third row.

The second and the obtained third rows are then used as starting rows to generate the new matrix H_2 and the block elements in the fourth row. The second and the obtained fourth rows are used to generate the new matrix K_3 and the fifth row. Also, the fourth row and the obtained fifth row are used to generate the new matrix H_4 and the sixth row.

Repeating the processes of determining K_3 and H_4 and the corresponding rows to the $2n+1$ row yields the complete array. It is noticed that the array exists if rank $D_{i,1} = m$ for $i = 1, 2, \ldots$, and also that the block-Routh array with the block-Routh algorithm is different from the matrix-Routh array with the matrix-Routh algorithm developed for matrix-continued fraction expansion.[12]

When any matrices $D_{i,1}$ are singular, the block-Routh array become a numerically ill-conditioned case. The original $T(s)$ will be modified according to applications in such a way that the block-Routh algorithm can still be applied. Various remedial methods will be suggested in the latter Sections.

Since the block algorithm in eqn. 9 shows the combinations of a repeated process and an alternately repeated process of the long divisions of two matrix polynomials, the algorithm can be expressed by the recurrence process as

(i)

$$
\begin{aligned}
T(s) &= D_2(s)D_1(s) \\
K_1 &= D_1(s)[sD_2(s)]^{-1} \quad \text{as} \quad s \to \infty \\
D_3(s) &= D_1(s) - sK_1D_2(s) \\
H_2 &= D_2(s)D_3(s)^{-1} \quad \text{as} \quad s \to \infty \\
D_4(s) &= D_2(s) - H_2D_3(s)
\end{aligned}
\qquad (10a)
$$

(ii)

$$\left.\begin{aligned} K_{i+1} &= D_i(s)[sD_{i+2}(s)]^{-1} \quad \text{as} \quad s\to\infty \\ D_{i+3}(s) &= D_i(s) - sK_{i+1}D_{i+2}(s) \\ H_{i+2} &= D_{i+2}(s)D_{i+3}(s)^{-1} \quad \text{as} \quad s\to\infty \\ D_{i+4}(s) &= D_{i+2}(s) - H_{i+2}D_{i+3}(s) \\ D_{2n+2}(s) &= 0_m \end{aligned}\right\}$$

$$i = 2,4,6,\ldots,2n-2 \qquad (10b)$$

The matrices K_i and H_i are called the block quotients, and are different from the matrix quotients[12] obtained from the matrix-Routh algorithm except for the first two block quotients.

Eqns. 10*a* and *b* can be combined and simplified as

$$\left.\begin{aligned} D_i(s)D_{i+2}(s)^{-1} &= Q_{i+1}(s) \\ &\quad - H_{i+2}^{-1}D_{i+4}(s)D_{i+2}(s)^{-1} \\ Q_{i+1}(s) &= H_{i+2}^{-1} + sK_{i+1} \end{aligned}\right\} \; i = 0,2,4,\ldots,2n-2$$

$$D_0(s) \triangleq D_1(s)$$

$$D_{2n+2}(s) \triangleq 0_m \qquad (10c)$$

Successively substituting eqn. 10*c* into $T(s)$ in eqn. 10*a* yields

$$\begin{aligned} T(s) &= D_2(s)D_1(s)^{-1} = [D_1(s)D_2(s)^{-1}]^{-1} \\ &= [Q_1(s) - H_2^{-1}D_4(s)D_2(s)^{-1}]^{-1} \\ &= [Q_1(s) - H_2^{-1}[D_2(s)D_4(s)^{-1}]^{-1}]^{-1} \\ &= [Q_1(s) - H_2^{-1}[Q_3(s) - H_4^{-1}[D_6(s)D_4(s)^{-1}]^{-1}]^{-1}]^{-1} \\ &= \ldots \\ &= [Q_1(s) - H_2^{-1}[Q_3(s) - H_4^{-1}[Q_5(s) \\ &\quad - H_6^{-1}[\ldots - H_{2n}^{-1}[Q_{2n-1}(s)]^{-1}]^{-1}]^{-1}]^{-1}]^{-1} \end{aligned}$$

where (10*d*)

$$Q_{i+1}(s) = sK_{i+1} + H_{i+2}^{-1} \quad i = 0,2,4,\ldots,2n-2$$

Eqn. 10*d* is a Stieltjes-type[23] matrix continued fraction. The counterpart of the scalar canonical form of eqn. 10*d* has been developed by Field and Owens.[29]

When the block quotients K_i and H_i are given and the original matrix polynomials are required, this is an inverse problem. The reverse recursive relations in eqn. 10 can be applied to determine the original matrix polynomials and listed as

(i)

$$D_{2n+2}^*(s) = 0_m \quad \text{and} \quad D_{2n+1}^*(s) = I_m \qquad (11a)$$

(ii)

$$\left.\begin{aligned} D_{i+2}^*(s) &= H_{i+2}D_{i+3}^*(s) + D_{i+4}^*(s) \\ D_i^*(s) &= sK_{i+1}D_{i+2}^*(s) + D_{i+3}^*(s) \\ D_{i+1}^*(s) &= H_i^{-1}[D_i^*(s) - D_{i+2}^*(s)] \end{aligned}\right\}$$

$$i = 2n-2, 2n-4, \ldots, 2 \qquad (11b)$$

(iii)

$$D_1^*(s) = sK_1D_2^*(s) + D_3^*(s) \qquad (11c)$$

The desired $D_2(s)$ and $D_1(s)$ in eqn. 2 become

$$D_2(s) = D_2^*(s)\left[\left(\prod_{i=1}^{n} K_{2i-1}\right)H_{2n}\right]^{-1} \qquad (11d)$$

$$D_1(s) = D_1^*(s)\left[\left(\prod_{i=1}^{n} K_{2i-1}\right)H_{2n}\right]^{-1} \qquad (11e)$$

and

$$T(s) = D_2^*(s)D_1^*(s)^{-1} = D_2(s)D_1(s)^{-1} \qquad (11f)$$

The inverse block-Routh algorithm of eqn. 9 is

(i)

$$D_{2n+2,j}^* = 0_m \qquad j = 1,2,\ldots$$

$$D_{2n+1,1}^* = I_m, D_{2n+1,j}^* = 0_m \qquad j = 2,3,\ldots \qquad (11g)$$

(ii)

$$\left.\begin{aligned} D_{i+2,1}^* &= H_{i+2}D_{i+3,1}^* \\ D_{i+2,j+1}^* &= D_{i+4,j}^* + H_{i+2}D_{i+3,j+1}^* \\ D_{i,1}^* &= K_{i+1}D_{i+2,1}^* \\ D_{i,j+1}^* &= D_{i+3,j}^* + K_{i+1}D_{i+2,j+1}^* \\ D_{i+1,j+1}^* &= H_i^{-1}[D_{i,j+1}^* - D_{i+2,j}^*] \end{aligned}\right\}$$

$$j = 1,2,3,\ldots \qquad i = 2n-2, 2n-4, \ldots, 2 \qquad (11h)$$

(iii)

$$\left.\begin{aligned} D_{1,1}^* &= K_1D_{21}^* \\ D_{1,j}^* &= D_{3,j}^* + K_1D_{2,j+1}^* \\ D_{1,j} &= D_{1,j}^*\left[\left(\prod_{i=1}^{n} K_{2i-1}\right)H_{2n}\right]^{-1} \\ D_{2,j} &= D_{2,j}^*\left[\left(\prod_{i=1}^{n} K_{2i-1}\right)H_{2n}\right]^{-1} \end{aligned}\right\} \quad j = 1,2,3,\ldots \qquad (11i)$$

From eqns. 9 and 11 we observe that the block-Routh algorithm involves only real matrices generated from the matrix coefficients in the numerator and denominator matrix polynomials of a matrix transfer function. Also we notice that only one block transformation is required to transform a block-state equation in eqn. 3 to a block-tridiagonal matrix form in eqn. 8. We believe that the proposed block-state equation in eqn. 8 and the block transformation in eqn. 7 are new.

3 Minimal realisations of matrix transfer functions

In the analysis and synthesis of the matrix transfer function of a multivariable system, the primary concerns are the internal structure and stability of the system. When $D_2(s)$ and $D_1(s)$ in $T(s) = D_2(s)D_1(s)^{-1}$ are coprime, the realisation of the $T(s)$ is minimal and the stability of the system can be determined from the scalar characteristic polynomial, $\det[D_1(s)]$. The minimal realisation is significant because the minimum number of integrators can set up an analogue or digital simulation of a matrix transfer function, and also more information about the internal structure of the

system can be obtained than from the original formulation. In the frequency domain, a matrix transfer function has been realised using various Cauer forms of matrix continued fractions.[12-14] When the nth- and the $(n-1)$th-degree matrix polynomials, $D_1(s)$ and $D_2(s)$, are arranged into the first two rows of the block-Routh array as shown in eqn. 9, the matrix-Routh algorithm[12] has been applied to determine $2n$ $m \times m$ matrix quotients H_i^*. It has been shown[12] that the state space equation with $nm \times nm$ system matrix, which is constructed using $2n$ matrix quotients H_i^*, is a minimal realisation of the $T(s)$ in the first Cauer matrix form. As a result, the $D_2(s)$ and $D_1(s)$ are right coprime. Using the same block elements in the first two rows and applying the proposed block-Routh algorithm in eqn. 9 results in the same number $(2n)$ of block quotients K_i and H_i. Therefore, the block-state equation constructed using the same number $(2n)$ of block quotients with dimensions $m \times m$ is a minimal realisation of the same $T(s)$. As a result, the block-state equation in eqn. 8 is completely controllable and observable, and the $D_2(s)$ and $D_1(s)$ are right coprime. It is noticed that only $K_1 = H_1^*$, $H_2 = H_2^*$ and other quotients are different.

An alternative way to show that the block-state equation in eqn. 8 is a minimal realisation of $T(s)$ can be described as follows: writing the controllability matrix (T_c) and the observability matrix (T_0)[27] in Kalman form result in

$$T_c = [E_1, \; G_1E_1, \; G_1^2E_1, \; \ldots, \; G_1^{n-1}E_1]$$

$$= \begin{bmatrix} 0_m & 0_m & \ldots & (K_1K_3 \ldots K_{2n-1})^{-1} \\ \cdot & \cdot & \ldots & \cdot \\ 0_m & (K_1K_3)^{-1} & \ldots & x \\ K_1^{-1} & x & \ldots & x \end{bmatrix}$$

and

$$T_0 = [F_1', \; (F_1G_1)', \; (F_1G_1^2)', \; \ldots, \; (F_1G_1^{n-1})']$$

$$= \begin{bmatrix} 0_m & 0_m & \ldots & ((H_{2n}K_{2n-1} \ldots H_2K_1)')^{-1} \\ \cdot & \cdot & \ldots & \cdot \\ 0_m & ((H_2K_1)')^{-1} & \ldots & x \\ I_m & x & \ldots & x \end{bmatrix}$$

where the xs denote unspecified blocks.

Both matrices T_c and T_0 are block triangular matrices. From the cross-diagonal block elements in T_c, we find that if rank $K_i = m$, then rank $T_c = nm$ and the system is completely controllable. Also, from T_0 we find that if rank $K_i = m$ and rank $H_i = m$, then rank $T_0 = nm$ and the system is completely observable. Thus, we can conclude that the system described in eqn. 8 is completely controllable and observable and it is a minimal realisation of $T(s)$ with minimal dimension nm if rank $K_i = m$ and rank $H_i = m$.

From the above conclusion we are now able to determine the condition that the block-Routh array exists. The necessary (but insufficient) condition for the existence of the block-Routh array is that the $q(= k/m)$ has to be an integer, where k is the rank of $T(s)$, (which can be determined from the Hankel matrix[15] or from the Gilbert's theorem,[24] and m is the input-output number. The sufficient condition is not added into the necessary condition because, for a rare case, some $D_{i,1}$ in the block-Routh array may be ill conditioned matrices even if q is an integer. If this is the case, the $T(s)$ is decomposed into two parallel subsystems $T_1(s)$ and $T_2(s)$ by using partial-fraction expansion. The rank of $T_1(s)$ is chosen as $m(q-l)$, and that of $T_2(s)$ as lm, where l is a proper integer. The proposed method can be applied to determine the minimal realisation of each $T_1(s)$ and $T_2(s)$. ~~Because the controllability and observability of a system are invariant[25] under such that rank $T_1(s) = mq$ and rank $T_2(s) = r$. Thus $T_1(s)$ the two parallel subsystems is the minimal realisation of the $T(s)$.~~

When the ratio k/m is not an integer, or $k = mq + r$, the two parallel subsystems $T_1(s)$ and $T_2(s)$ shall be chosen such that rank $T_1(s) = mq$ and rank $T_2(s) = r$. Thus $T_1(s)$ can be realised by the proposed method and $T_2(s)$ can be realised by using any other methods,[15] such as the Gilbert's method.[24] The composite state equation of the two parallel subsystems is the minimal realisation of the $T(s)$.

In order to obtain the minimal realisation of a general matrix transfer function $T(s)$ that contains not coprime matrix polynomials, we consider $T(s)$ as

$$\begin{aligned} T(s) &= D_2(s)D_1(s)^{-1} \\ &= [D_{21}s^{n-1} + D_{22}s^{n-2} + \ldots + D_{2,n}] \\ &\quad \times [D_{11}s^n + D_{12}s^{n-1} + \ldots + D_{1,n+1}]^{-1} \\ &= P_2(s)C(s)[P_1(s)C(s)]^{-1} \\ &= P_2(s)P_1(s)^{-1} \\ &= [P_{21}s^{r-1} + P_{22}s^{r-2} + \ldots + P_{2,r}] \\ &\quad \times [P_{11}s^r + P_{12}s^{r-1} + \ldots + P_{1,r+1}]^{-1} \end{aligned} \tag{12a}$$

where

$$C(s) = C_{n+1-r}s^{n-r} + C_{n-r}s^{n-r-1} + \ldots + C_1 \tag{12b}$$

$C(s)$ is a common matrix polynomial, and $D_2(s)$ and $D_1(s)$ are not coprime. The realisation of $T(s)$ using $D_2(s)$ and $D_1(s)$ is not minimal, and the stability cannot be determined from the scalar polynomial det $[D_1(s)]$ or det $[sI - A]$ in eqn. 3.

The $C(s)$ can be extracted from the block-Routh array in eqn. 9 as follows: when $n = r$, in eqn. 12, the block-Routh array in eqn. 9 terminates normally, and $C(s) = I_m$. When $n > r$, the array will terminate prematurely. The $C(s)$ can be obtained from the last nonvanishing row in eqn. 9. For example, if $D_{2n,1} = 0_m$, then $C(s) = D_{2n-1,1}s + D_{2n-1,2}$. This can be verified from the following: when $2r$ block quotients K_i and H_i are available, the matrix coefficients $P_{1,i}$ and $P_{2,i}$ in eqn. 12 can be determined from the reverse process of the algorithm in eqn. 11, or

(i)

$$P_{2r+2}(s) = 0_m, P_{2r+1}(s) \triangleq I_m \tag{13a}$$

(ii)

$$\left.\begin{aligned} P_{i+2}(s) &= H_{i+2}P_{i+3}(s) + P_{i+4}(s) \\ P_i(s) &= sK_{i+1}P_{i+2}(s) + P_{i+3}(s) \\ P_{i+1}(s) &= H_i^{-1}[P_i(s) - P_{i+2}(s)] \end{aligned}\right\}$$

$$i = 2r-2, 2r-4, \ldots, 2 \tag{13b}$$

(iii)

$$P_1(s) = sK_1P_2(s) + P_3(s) \tag{13c}$$

If the array terminates prematurely, the $D_2(s)$ and $D_1(s)$ can be obtained in the same fashion as

(i)

$$D_{2r+2}(s) = 0_m, \qquad D_{2r+1}(s) = I_m C(s) \tag{14a}$$

(ii)

$$\left.\begin{aligned} D_{i+2}(s) &= H_{i+2}D_{i+3}(s) + D_{i+4}(s) \\ D_i(s) &= sK_{i+1}D_{i+2}(s) + D_{i+3}(s) \\ D_{i+1}(s) &= H_i^{-1}[D_i(s) - D_{i+2}(s)] \end{aligned}\right\}$$

$$i = 2r-2, 2r-4, \ldots, 2 \tag{14b}$$

(iii)

$$D_1(s) = sK_1D_2(s) + D_3(s) \tag{14c}$$

Substituting eqn. 13 into eqn. 14 yields

$$D_{2r+1}(s) = P_{2r+1}(s)C(s) \tag{15a}$$

and

$$D_i(s) = P_i(s)C(s), \qquad i = 2r-2, 2r-4, \ldots, 2 \tag{15b}$$

Finally, we have

$$D_2(s) = P_2(s)C(s) \tag{15c}$$

$$D_1(s) = P_1(s)C(s) \tag{15d}$$

Therefore $C(s)$ is the common matrix polynomial of $D_2(s)$ and $D_1(s)$ which can be determined from the nonvanishing row in the block-Routh array. The $P_2(s)$ and $P_1(s)$ are right coprime and the realisation of $T(s)$ using $P_2(s)$ and $P_1(s)$ is minimal. Thus, $C(s)$ is the greatest common matrix polynominal.[27, 31] The corresponding block diagram of the minimal realisation of the $T(s)$ in eqn. 3 is shown in Fig. 1.

An alternative method, which uses the matrix-Routh array but not the block-Routh array, has been proposed by Shieh, *et al.*[25] to extract the common matrix polynominal of the two matrix polynominals that are not coprime.

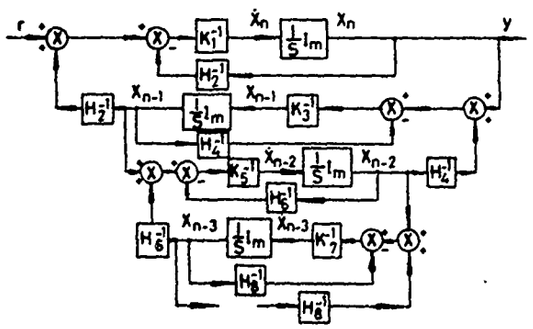


Fig. 1 *Block diagram representation of the block-state equation in eqn. 8*

4 State-space realisation of driving-point impedance matrix

When the input vector $R(s)$ and the output vector $Y(s)$ in eqn. 2 are observed as the input-current vector and the output-voltage vector at the same port, the matrix transfer function $T(s)$ becomes the driving-point impedance matrix $Z(s)$. The realisation of this positive real impedance matrix via state-space approaches has been successfully developed by Anderson and Vongpanitlerd.[15] Multiwinding transformers are mainly utilised for the realisations. Recently, Takahashi *et al.*[6] have proposed an elegant method for transfer-function realisations without using integrators. However, they deal with only one-port networks. In this paper, their method will be extended to realise the $Z(s)$ using the newly developed block-state equation in the block-tridiagonal form.

The $Z(s)$ of interest in this section is an RC driving-point impedance matrix. The multi-port RC ladder network obtained from the $Z(s)$ will be used as the plant in synthesising the general matrix transfer function. The block-state equation of a multi-port RC ladder network in Fig. 2 is

$$\dot{v} = G_0 v + E_0 r \tag{16a}$$

$$y = F_0 v \tag{16b}$$

where

$$G_0 = \begin{bmatrix} -C_1^{-1/2}R_2^{-1}C_1^{-1/2} & C_1^{-1/2}R_2^{-1}C_2^{-1/2} & 0_m & \cdots & 0_m & 0_m \\ C_2^{-1/2}R_2^{-1}C_1^{-1/2} & -C_2^{-1/2}(R_2^{-1}+R_3^{-1})C_2^{-1/2} & C_2^{-1/2}R_3^{-1}C_3^{-1/2} & \cdots & 0_m & 0_m \\ 0_m & C_3^{-1/2}R_3^{-1}C_2^{-1/2} & -C_3^{-1/2}(R_3^{-1}+R_4^{-1})C_3^{-1/2} & \cdots & 0_m & 0_m \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0_m & 0_m & 0_m & \cdots & C_{n-1}^{-1/2}R_n^{-1}C_n^{-1/2} & -C_n^{-1/2}(R_n^{-1}+R_{n+1}^{-1})C_n^{-1/2} \end{bmatrix},$$

$$E_0 = \begin{bmatrix} C_1^{-1/2} \\ 0_m \\ 0_m \\ \vdots \\ 0_m \\ 0_m \end{bmatrix}, \qquad v = \begin{bmatrix} C_1^{1/2}v_1 \\ C_2^{1/2}v_2 \\ C_3^{1/2}v_3 \\ \vdots \\ C_{n-1}^{1/2}v_{n-1} \\ C_n^{1/2}v_n \end{bmatrix}$$

$$F_0 = [C_1^{-1/2} \quad 0_m \quad 0_m \quad 0_m \quad \cdots \quad 0_m \quad 0_m]$$

Note that R_i and C_i are real, symmetric and positive-definite matrices, so that

$$C_{i-1}^{-1/2}R_i^{-1}C_i^{-1/2} = (C_{i-1}^{-1/2}R_i^{-1}C_i^{-1/2})' = C_i^{-1/2}R_i^{-1}C_{i-1}^{-1/2} \quad (17a)$$

$$C_1^{-1/2}R_2^{-1}C_1^{-1/2} = (C_1^{-1/2}R_2^{-1}C_1^{-1/2})' \quad (17b)$$

and

$$C_i^{-1/2}(R_i^{-1} + R_{i+1}^{-1})C_i^{-1/2} = [C_i^{-1/2}(R_i^{-1} + R_{i+1}^{-1})C_i^{-1/2}]' \quad (17c)$$

In other words, the system matrix G_0 is a symmetric matrix. In order to match the block elements of the block-state equations in eqns. 8 and 16, we perform the following block transformation on the block-state equation in eqn. 8:

$$z = T_2^{-1}q \quad (18)$$

where

$$T_2 = \begin{bmatrix} 0_m & 0_m & . & 0_m & L_1 \\ 0_m & 0_m & . & L_2 & 0_m \\ . & . & . & . & . \\ 0_m & L_{n-1} & . & 0_m & 0_m \\ L_n & 0_m & . & 0_m & 0_m \end{bmatrix}$$

The new block-state equation becomes

$$\dot{q} = T_2G_1T_2^{-1}q + T_2E_1r = G_2q + E_2r \quad (19a)$$

$$y = F_1T_2^{-1}q = F_2q \quad (19b)$$

where

$$G_2 = \begin{bmatrix} -M_1 & N_1 & 0_m & 0_m & . & 0_m & 0_m \\ N_1 & -M_2 & N_2 & 0_m & . & 0_m & 0_m \\ 0_m & N_2 & -M_3 & N_3 & . & 0_m & 0_m \\ . & . & . & . & . & . & . \\ 0_m & 0_m & 0_m & 0_m & . & -M_{n-1} & N_{n-1} \\ 0_m & 0_m & 0_m & 0_m & . & N_{n-1} & -M_n \end{bmatrix},$$

$$E_2 = \begin{bmatrix} L_1K_1^{-1} \\ 0_m \\ 0_m \\ . \\ 0_m \\ 0_m \end{bmatrix}$$

$$F_2 = [L_1^{-1} \quad 0_m \quad 0_m \quad 0_m \quad . \quad 0_m \quad 0_m] .$$

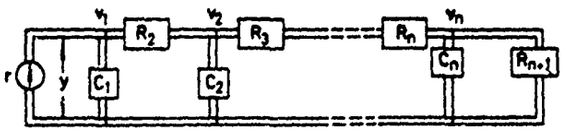


Fig. 2 *Multiport RC ladder network*

The recursive relations of the block elements L_i, M_i, N_i, K_i and H_i are as follows:

$$L_1 = \text{any nonsingular matrix}$$

$$M_i = L_i(H_{2i}K_{2i-1})^{-1}L_i^{-1} \qquad i = 1, 2, \ldots, n$$

$$N_i = [L_i(K_{2i+1}H_{2i}K_{2i-1})^{-1}L_i^{-1}]^{1/2} \qquad i = 1, 2, \ldots, n-1$$

$$L_{i+1} = N_iL_iK_{2i+1} \qquad i = 1, 2, \ldots, n-1 \quad (20)$$

The square root of a matrix can be determined using the method suggested by Frame.[16] Comparing the respective E_2 and F_2 in eqn. 19 with E_0 and F_0 in eqn. 16 yields $L_1 = K_1^{1/2} = C_1^{1/2}$. When the block elements K_i, N_i and M_i are real, symmetric and positive definite, we can solve the R_i and C_i by matching the block elements in the system matrix G_0 in eqn. 16 and that of G_2 in eqn. 19 starting from the block elements M_1, N_1, M_2 and $N_2, \ldots$.

The $K_1 (= C_1)$ is chosen as a real, symmetric and positive-definite matrix in order that the synthesis of this capacitor matrix C_1 can be performed by using multiwinding transformers. Also, the block elements N_i and M_i are restricted as real, symmetric and positive-definite matrices so that G_2 in eqn. 19 has real eigenvalues, owing to the symmetric property of this matrix. In the following Section, it will be further shown that the system in eqn. 19 is asymptotically stable. As a result, the real eigenvalues of G_2 are negative real.

Thus the RC driving-point impedance matrix can be synthesised using multiport RC network. It is noticed that the passive RC structures are different from that of Cauer's first form. Thus, extending the ideas of Takahashi *et al.*[6] and the newly developed block-tridiagonal form, a matrix transfer function may be synthesised via block-state space approaches without using integrators.

When the block-Routh array of an RC driving-point impedance matrix $Z(s)$ becomes an ill conditioned case, the proposed remedial methods in Section 3 can be applied to overcome the difficulty. The synthesised multiport RC networks of the decomposed subsystems $Z_1(s)$ and $Z_2(s)$ are connected in cascade[14] because the $Z(s) (= Z_1(s) + Z_2(s))$ is a driving-point impedance matrix rather than a transfer-function matrix.

When the driving-point impedance function of a one-port network is of interest, the linear transformation T_2 in eqn. 18, and the system matrix G_2 in eqn. 19, can be expressed in terms of scalar quotients k_i and h_i (instead of K_i and H_i) obtained from the modified Routh array and Routh algorithm in eqn. 9. The linear transformation matrix T_2 is

$$T_2 = \begin{bmatrix} 0 & 0 & . \\ 0 & 0 & . \\ 0 & 0 & . \\ . & . & . \\ \sqrt{\dfrac{k_{2n-1}}{h_0h_2 \ldots h_{2n-2}}} & 0 & . \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & \sqrt{\frac{k_1}{h_0}} \\ 0 & \sqrt{\frac{k_3}{h_0 h_2}} & 0 \\ \sqrt{\frac{k_5}{h_0 h_2 h_4}} & 0 & 0 \\ \cdot & \cdot & \cdot \\ 0 & 0 & 0 \end{bmatrix} \quad (21)$$

where $h_0 = 1$. The system matrix G_2 becomes

$$G_2 = \begin{bmatrix} -\frac{1}{k_1 h_2} & \frac{1}{\sqrt{k_1 h_2 k_3}} & 0 \\ \frac{1}{\sqrt{k_1 h_2 k_3}} & -\frac{1}{k_3 h_4} & \frac{1}{\sqrt{k_3 h_4 k_5}} \\ 0 & \frac{1}{\sqrt{k_3 h_4 k_5}} & -\frac{1}{k_5 h_6} \\ \cdot & \cdot & \cdot \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{matrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \cdot & \cdot \\ \frac{1}{\sqrt{k_{2n-3} h_{2n-2} k_{2n-1}}} & -\frac{1}{k_{2n-1} h_{2n}} \end{matrix} \quad (22a)$$

The input vector E_2 and the output vector F_2 are

$$E_2 = \left[\frac{1}{\sqrt{k_1}} \quad 0 \quad 0 \quad . \quad 0 \quad 0\right]' \quad (22b)$$

$$F_2 = \left[\frac{1}{\sqrt{k_1}} \quad 0 \quad 0 \quad . \quad 0 \quad 0\right] \quad (22c)$$

Note that the process to evaluate the elements in T_2 and G_2 involves only real numbers rather than complex numbers, as suggested by Takahashi *et al.*[6]

5 Stability of matrix transfer functions

When a multivariable system is represented by a matrix transfer function, $T(s) = D_2(s)D_1(s)^{-1}$, and the stability of the system is required to be determined, one often converts this matrix transfer function into a high-dimensional state equation in general co-ordinates, and determines the scalar characteristic equation. The stability of this system is then determined either by directly applying the Routh criterion[17] or application of Jury's inner theory[3] on the scalar characteristic polynominal. However, it is tedious to determine a scalar characteristic polynominal of a large-dimensional system. Furthermore, when $D_1(s)$ and $D_2(s)$ are not coprime, the scalar polynomial obtained is not the characteristic polynomial. Several authors have studied the stability of a multivariable system from the characteristic matrix polynomial. Papaconstantinou[18] suggested a recursive algorithm for indirectly determining the stability of a matrix polynomial. Anderson and Bitmead[19] determine the stability of a matrix polynomial by testing the lossless positive real of a rational transfer function matrix which is derived from $D_1(s)$. Denman[20] has also suggested a numerical method to determine block roots of a matrix polynomial which can be used to determine the stability of a matrix polynomial. Recently, Shieh *et al.*[2, 11] have partially extended the scalar Routh criterion to the matrix Routh criterion for testing the stability. All above methods have assumed that $D_1(s)$ is the characteristic matrix polynomial. In this paper, we develop a method to test the stability of a matrix transfer function in which $D_2(s)$ and $D_1(s)$ may not be coprime.

Performing the following block transformation:

$$q = T_3^{-1}\omega \quad (23a)$$

where

$$T_3 = \text{block diag.}\,(I_m, jI_m, I_m, jI_m, \ldots) \quad (23b)$$

on eqn. 19 gives

$$\dot{\omega} = T_3 G_2 T_3^{-1}\omega + T_3 E_2 r = G_3\omega + E_3 r \quad (24a)$$

$$y = F_2 T_3^{-1}\omega = F_3\omega \quad (24b)$$

where

$$G_3 = \begin{bmatrix} -M_1 & -jN_1 & . & 0_m & 0_m \\ jN_1 & -M_2 & . & 0_m & 0_m \\ \cdot & \cdot & . & \cdot & \cdot \\ 0_m & 0_m & . & -M_{n-1} & -jN_{n-1} \\ 0_m & 0_m & . & jN_{n-1} & -M_n \end{bmatrix},$$

$$E_3 = \begin{bmatrix} L_1 K_1^{-1} \\ 0_m \\ \cdot \\ 0_m \\ 0_m \end{bmatrix}$$

$$F_3 = [L_1^{-1} \quad 0_m \quad . \quad 0_m \quad 0_m]$$

and $j = \sqrt{-1}$.

Now, consider the following quadratic equation:

$$v = \omega' P \omega \quad (25a)$$

where

$$P = \text{block diag.}\,[I_m, \; I_m, \; . \; I_m, \; I_m] \quad (25b)$$

Since P is positive definite, v is positive definite. When $N_i = N_i'$ in eqn. 24, the derivative of v is

$$\dot{v} = \omega'[PG_3 + G_3'P]\omega = -\omega' Q \omega \quad (26a)$$

where

$$Q = 2 \times \text{block diag.}\,[\bar{M}_1, \bar{M}_2, \ldots, \bar{M}_n] \quad (26b)$$

and

$$\bar{M}_i = [M_i + M_i']/2 \qquad i = 1, 2, \ldots, n \quad (26c)$$

If $\bar{M}_i$ are real symmetric and positive-definite matrices, this implies that Q is positive definite and the v in eqn. 25 is a Lyapunov function. From Lyapunov theory[21] we can conclude that the system in eqn. 2 is asymptotically stable if

(i) $N_i = N_i'$ and M_i are real and positive definite

(ii) $det\ A_{n+1}$ and $det\ A_1$ in eqn. 2 have the same sign, and are nonzero.

It is noticed that N_i may be real or imaginary matrices; M_i may be nonsymmetric. A_{n+1} and A_1 are the matrix coefficients of the characteristic matrix polynomial $D_1(s)$ in eqn. 2, therefore the $det\ A_{n+1}$ is the leading coefficient of s^{nm} and $det\ A_1$ is the constant term in the scalar polynomial $det\ [D_1(s)]$. When $D_2(s)$ and $D_1(s)$ are coprime, the $det\ [D_1(s)]$ is the characteristic polynomial of the system. The necessary condition for the stable system states that $det\ A_{n+1}$ and $det\ A_1$ should have the same sign. Even if $D_2(s)$ and $D_1(s)$ are not coprime, we have the same block quotients K_i and H_i, or the block elements N_i and M_i; however, the $det\ A_{n+1}$ and $det\ A_1$ shall be replaced by $det\ P_{11}$ and $det\ P_{1,r+1}$, which can be obtained from the matrix coefficients of $D_1(s)$, $D_2(s)$ and $C(s)$ in eqn. 12 and expressed as

$$det\ P_{11} = det\ [A_{n+1}C_{n+1-r}^{-1}] \tag{27a}$$

$$det\ P_{1,r+1} = det\ [A_1C_1^{-1}] \tag{27b}$$

When rank $T(s)(=k)$ is not equal to mq, (or $k = mq + r$), the block-Routh array becomes an ill conditioned case. The $T(s)$ can be modified by adding another stable matrix, $T_3(s)(=K/(s+\alpha)$ where K is a constant matrix with rank $K = m - r$ and α is a positive value), to form a new transfer function matrix $T^*(s)$ such that rank $T^*(s) = m(q+1)$. Thus, the proposed method can be applied to determine the block-Routh array and the stability of the system. When $k = mq$ and the ill conditioned case occurs, the $T(s)$ is modified by multiplying a stable diagonal matrix with stable diagonal elements as $(s+\beta)/(s+\alpha)$, where β and α are positive values. Thus, the proposed method can be applied to determine the block-Routh array and the stability. It is noticed that the stability of a multivariable system is invariant under such modifications.

When the matrix transfer function is completely decoupled such that A_i and B_i in eqn. 2 are diagonal matrices, then the recursive algorithm in eqn. 20 can be further simplified as

$$M_i = (H_{2i}K_{2i-1})^{-1} \qquad i = 1, 2, \ldots, n \tag{28a}$$

$$N_i = (K_{2i+1}H_{2i}K_{2i-1})^{-1/2} \qquad i = 1, 2, \ldots, n-1 \tag{28b}$$

where H_i, K_i, M_i and N_i are diagonal matrices.

If $det\ P_{11}$ and $det\ P_{1,r+1}$ in eqn. 27 have the same sign and the pairs $\{H_{2i}\ K_{2i-1}\}$ are positive definite, then the system in eqn. 2 is asymptotically stable. Furthermore, if all K_i and H_i are positive definite, then the system is not only asymptotically stable, but also the poles and zeros of each transfer function $y_i(s)/R_i(s)$ interlace on the negative real axis of the s-plane. This can be verified as follows. Each transfer function $y_i(s)/R_i(s)$ can be considered as a driving-point impedance function. Comparing G_2 in eqn. 22a and the G_0 in eqn. 16, we can solve the positive values of R_i and C_i. The network realisation of the impedance function is an RC-type ladder network[22] as shown in Fig. 2. Therefore, the $y_i(s)/R_i(s)$ is not only a positive real function, but also the poles and zeros must alternate on the negative real axis in the s-plane. From above properties we are now able to synthesise a matrix transfer function, $T(s) = D_2(s)D_1(s)^{-1}$ in eqn. 2, without using integrators and multiwinding transformers. The steps are described as follows:

Step 1 Construct m sets of independent one-port RC ladder networks that contain any proper values of resistors and capacitors. As a result, R_i and C_i in Fig. 2 are positive-definite diagonal matrices.

Step 2 Match the block elements in eqn. 16 and those of M_i and N_i in eqn. 19 to determine the diagonal matrices H_i and K_i in eqn. 28 using R_i and C_i.

Step 3 Substitute the obtained H_i and K_i into eqn. 11 to determine the matrix polynomials $D_1^{**}(s)$ and $D_2^{**}(s)$ having diagonal matrix coefficients.

Step 4 Subtract the matrix coefficients of the same power in $D_1(s)$ and $D_1^{**}(s)$ to obtain the feedback block gains and also the matrix coefficients in $D_2(s)$ and $D_2^{**}(s)$ to determine the feed-forward block gains in the phase-variable block co-ordinates as shown in eqn. 3.

Step 5 Transform the above block gains from the phase-variable block co-ordinates to the tridiagonal block co-ordinates in eqn. 19 using the block transformations in eqns. 6 and 18.

Step 6 Sum up the feedback block gains using one block summer, and the feed-forward block gains using another block summer.

Thus, a matrix transfer function can be synthesised using a state-space approach.

6 Illustrative examples

Example 1

To illustrate the processes, we determine a block-transformation matrix, a block-state equation in a block-tridiagonal form, the stability, and the state-space realisation of the following driving-point impedance matrix that is represented by a matrix transfer function:

$$Y(s) = T(s)R(s) \tag{29}$$

where

$$\begin{aligned} T(s) &= D_2(s)D_1(s)^{-1} \\ &= [D_{21}s + D_{22}][D_{11}s^2 + D_{12}s + D_{13}]^{-1} \\ &= \left[\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}s + \begin{pmatrix} 24 & 10 \\ 22 & 10 \end{pmatrix}\right] \\ &\quad \times \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}s^2 + \begin{pmatrix} 165 & 68 \\ 68 & 35 \end{pmatrix}s + \begin{pmatrix} 14 & 6 \\ 18 & 8 \end{pmatrix}\right]^{-1} \end{aligned}$$

Note that $D_2(s)$ and $D_1(s)$ are not symmetric matrix polynomials. The procedures can be shown in the following steps:

Step 1 Construct the block-state equation in the block companion form

$$\dot{x} = Ax + Br \tag{30a}$$

$$y = Cx$$

where

$$A = \begin{bmatrix} 0_2 & I_2 \\ -D_{13} & -D_{12} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ -\begin{pmatrix} 14 & 6 \\ 18 & 8 \end{pmatrix} & -\begin{pmatrix} 165 & 68 \\ 68 & 35 \end{pmatrix} \end{bmatrix},$$

$$B = \begin{bmatrix} 0_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{bmatrix}$$

$$C = [D_{22} \quad D_{21}] = \left[\begin{pmatrix} 24 & 10 \\ 22 & 10 \end{pmatrix} \quad \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\right]$$

Step 2 Construct the block-Routh array to determine the block quotients. The block-Routh array is

$$D_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad D_{12} = \begin{pmatrix} 165 & 68 \\ 68 & 35 \end{pmatrix}$$

$$D_{13} = \begin{pmatrix} 14 & 6 \\ 18 & 8 \end{pmatrix}$$

$$D_{21} = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \qquad D_{22} = \begin{pmatrix} 24 & 10 \\ 22 & 10 \end{pmatrix}$$

$$D_{31} = D_{12} - K_1 D_{22} = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}$$

$$D_{32} = D_{13} = \begin{pmatrix} 14 & 6 \\ 18 & 8 \end{pmatrix}$$

$$D_{41} = D_{22} - H_2 D_{32} = \begin{pmatrix} 10 & 4 \\ 4 & 2 \end{pmatrix}$$

$$D_{51} = D_{22} = \begin{pmatrix} 24 & 10 \\ 22 & 10 \end{pmatrix} \tag{30b}$$

where

$$K_1 = D_{11} D_{21}^{-1} = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

$$H_2 = D_{21} D_{31}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$K_3 = D_{21} D_{41}^{-1} = \begin{pmatrix} 2{\cdot}5 & -6 \\ -6 & 14{\cdot}5 \end{pmatrix}$$

$$H_4 = D_{41} D_{51}^{-1} = \begin{pmatrix} 0{\cdot}6 & -0{\cdot}2 \\ -0{\cdot}2 & 0{\cdot}4 \end{pmatrix}$$

Since we have $2n(=4)$ block quotients and the block-Routh array terminates normally, $D_2(s)$ and $D_1(s)$ are right coprime.

Step 3 Establish the block transformation T_1 in eqn. 7

$$x = T_1^{-1} z \tag{30c}$$

where

$$T_1^{-1} = \begin{bmatrix} D_{41} & 0_2 \\ D_{22} & D_{21} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 10 & 4 \\ 4 & 2 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 24 & 10 \\ 22 & 10 \end{pmatrix} & \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \end{bmatrix}$$

Step 4 The required block-state equation in eqn. 8 is

$$\dot{z} = G_1 z + E_1 r \qquad y = F_1 z \tag{30d}$$

where

$$G_1 = \begin{bmatrix} -(H_4 K_3)^{-1} & K_3^{-1} \\ (H_2 K_1)^{-1} & -(H_2 K_1)^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} -\begin{pmatrix} 140 & 130 \\ 58 & 54 \end{pmatrix} & \begin{pmatrix} 58 & 24 \\ 24 & 10 \end{pmatrix} \\ \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} & -\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \end{bmatrix}$$

$$E_1 = \begin{bmatrix} 0_2 \\ K_1^{-1} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \end{bmatrix}$$

$$F_1 = [0_2 \quad I_2] = \left[\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

Step 5 Evaluate M_t and N_t in eqn. 20 to determine the stability. For the use of multiport network synthesis, we choose

$$L_1 = K_1^{1/2} = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}^{1/2} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \tag{30e}$$

Thus, we solve M_t, N_t and L_t as

$$M_1 = L_1 (H_2 K_1)^{-1} L_1^{-1} = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}$$

$$N_1 = [L_1 (K_3 H_2 K_1)^{-1} L_1^{-1}]^{1/2} = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}$$

$$L_2 = N_1 L_1 K_3 = \begin{pmatrix} -0{\cdot}5 & 1{\cdot}5 \\ 1{\cdot}5 & -3{\cdot}5 \end{pmatrix}$$

$$M_2 = L_2 (H_4 K_3)^{-1} L_2^{-1} = \begin{pmatrix} 167 & 67 \\ 67 & 27 \end{pmatrix}$$

Since $det\, D_{11}(=1) > 0$, $det\, D_{13}(=4) > 0$, $N_1 = N_1'$, and M_1 and M_2 are real, symmetric and positive definite, the system is asymptotically stable. It is interesting to note that the poles of this multivariable system or the roots of $det\,[D_1(s)]$ are $s_1 = -0{\cdot}02739$, $s_2 = -0{\cdot}127864$,

$s_3 = -5{\cdot}88774$ and $s_4 = -193{\cdot}957$. The transmission zeros of this multivariable system or the roots of *det* $[D_2(s)]$ are $s_1 = -0{\cdot}10315$ and $s_2 = -193{\cdot}89685$.

Step 6 Compare the respective E_2 and G_2 in eqn. 19 and the E_0 and G_0 in eqn. 16 to solve R_l and C_l.

$$G_0 = \begin{bmatrix} -C_1^{-1/2}R_2^{-1}C_1^{-1/2} & C_1^{-1/2}R_2^{-1}C_2^{-1/2} \\ C_2^{-1/2}R_2^{-1}C_1^{-1/2} & -C_2^{-1/2}(R_2^{-1}+R_3^{-1})C_2^{-1/2} \end{bmatrix}$$

$$= \begin{bmatrix} -M_1 & N_1 \\ N_1 & -M_2 \end{bmatrix} = G_2 \qquad (30f)$$

$$E_0 = \begin{bmatrix} C_1^{-1/2} \\ 0_2 \end{bmatrix} = \begin{bmatrix} L_1 K_1^{-1} \\ 0_2 \end{bmatrix} = E_2$$

From eqn. 30*f*, we have

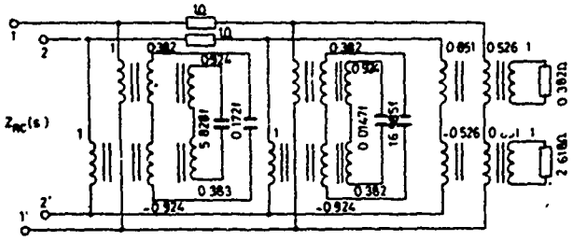
$$L_1 = K_1^{1/2} = C_1^{1/2} \rightarrow C_1 = K_1 = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

$$C_1^{-1/2}R_2^{-1}C_1^{-1/2} = M_1 \rightarrow R_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$C_1^{-1/2}R_2^{-1}C_2^{-1/2} = N_1 \rightarrow C_2 = \begin{pmatrix} 2{\cdot}5 & -6 \\ -6 & 14{\cdot}5 \end{pmatrix}$$

$$C_2^{-1/2}(R_2^{-1}+R_3^{-1})C_2^{-1/2} = M_2 \rightarrow R_3 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

The structure of the multiport network is shown in Fig. 2. Multiwinding transformers may be used to realise the R_l



$$Y_1 = T_1[sC_1]T_1' = sD_1$$

$$T_1 = \begin{bmatrix} 0{\cdot}9238 & 0{\cdot}3826 \\ 0{\cdot}3826 & -0{\cdot}9238 \end{bmatrix}, C_1 = \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}, D_1 = \begin{bmatrix} 5{\cdot}8284 & 0 \\ 0 & 0{\cdot}1716 \end{bmatrix}$$

$$Z_2 = R_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$Y_3 = T_3[sC_3]T_3 = sD_3$$

$$T_3 = \begin{bmatrix} 0{\cdot}9239 & 0{\cdot}3827 \\ 0{\cdot}3827 & -0{\cdot}9239 \end{bmatrix}, C_3 = \begin{bmatrix} 2{\cdot}5 & -6 \\ -6 & 14{\cdot}5 \end{bmatrix}, D_3 = \begin{bmatrix} 0{\cdot}0147 & 0 \\ 0 & 16{\cdot}9852 \end{bmatrix}$$

$$Z_4 = T_3RT_3 = D_3$$

$$T_3 = \begin{bmatrix} 0{\cdot}5257 & 0{\cdot}8507 \\ 0{\cdot}8507 & -0{\cdot}5257 \end{bmatrix}, R_3 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, D_3 = \begin{bmatrix} 0{\cdot}382 & 0 \\ 0 & 2{\cdot}618 \end{bmatrix}$$

Fig. 3 *Realisation of $Z_{RC}(s)$ in example 1*

and C_l. The network configuration is shown in Fig. 3. For this *RC* driving-point impedance matrix, the matrix Cauchy index due to Bitmead and Anderson[28] can be applied to determine the number of negative real roots and the stability.

It is noticed that, in determining the stability of a transfer function matrix, the N_l may be real or imaginary matrices. For example, if we use the same K_1 (defined as K_1^*) and H_2 (defined as H_2^*) in eqn. 30*b*, and assign $K_3^* = -K_3$ and $H_4^* = -H_4$, we have the same $L_1^*(=L_1)$, $M_1^*(=M_1)$, and $M_2^*(=M_2)$ as shown in eqn. 30*e*. Also, we have the imaginary matrices $L_2^*(=jL_2)$ and $N_1^*(=jN_1)$. Substituting K_l^* and H_l^* into eqn. 11 gives the transfer function matrix $T^*(s)$ as

$$T^*(s) = D_2^*(s)D_1^*(s)^{-1}$$

$$= [D_{21}^*s + D_{22}^*][D_{11}^*s^2 + D_{12}^*s + D_{13}^*]^{-1}$$

$$= \left[\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}s + \begin{pmatrix} 24 & 10 \\ 22 & 10 \end{pmatrix}\right]$$

$$\times \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}s^2 + \begin{pmatrix} 165 & 68 \\ 68 & 35 \end{pmatrix}s + \begin{pmatrix} 34 & 14 \\ 26 & 12 \end{pmatrix}\right]^{-1} \qquad (30g)$$

Since

$$det\ D_{11}^*(=1) > 0, det\ D_{13}^*(=44) > 0,$$

$$N_1^* = N_1^{*\prime}(=jN_1 = jN_1')$$

and both $M_1^*(=M_1)$ and $M_2^*(=M_2)$ are real and positive definite, then $T^*(s)$ is asymptotically stable. It might be interesting to know the distribution of the roots of *det* $D_1^*(s)$. The roots are

$$s_1 = -0{\cdot}197735 + j0{\cdot}160018,$$

$$s_2 = -0{\cdot}197735 - j0{\cdot}160018,$$

$$s_3 = -5{\cdot}767861 \quad \text{and} \quad s_4 = -193{\cdot}8367.$$

From the roots we observe that there exists a pair of complex poles in $T^*(s)$ and $T^*(s)$ is not an *RC* positive real matrix. Of course, the system matrix G_2 in eqn. 19 is a symmetric but not real matrix.

Example 2

Determine the pair of coprime matrix polynomials ($P_1(s)$ and $P_2(s)$), the common matrix polynomial $C(s)$ in eqn. 12, and the stability of the following matrix transfer function:

$$Y(s) = T(s)R(s) \qquad (31)$$

where

$$T(s) = D_2(s)D_1(s)^{-1} = [P_2(s)C(s)][P_1(s)C(s)]^{-1} = [D_{21}s^2 + D_{22}s + D_{23}][D_{11}s^3 + D_{12}s^2 + D_{13}s + D_{14}]^{-1}$$

$$= \left[\begin{pmatrix} 2 & -1 \\ -4 & 3 \end{pmatrix}s^2 + \begin{pmatrix} -333 & 338 \\ -141 & 147 \end{pmatrix}s + \begin{pmatrix} 338 & 172 \\ 146 & 74 \end{pmatrix}\right] \times$$

$$\left[\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}s^3 + \begin{pmatrix} -1945 & 1983 \\ -811 & 826 \end{pmatrix}s^2 + \begin{pmatrix} 1815 & 1177 \\ 753 & 492 \end{pmatrix}s + \begin{pmatrix} 169 & 86 \\ 73 & 37 \end{pmatrix}\right]^{-1}$$

$n = 3$ and $m = 2$.

The block-Routh array is

$$D_{11} = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \quad D_{12} = \begin{pmatrix} -1945 & 1983 \\ -811 & 826 \end{pmatrix}$$

$$D_{13} = \begin{pmatrix} 1815 & 1177 \\ 753 & 492 \end{pmatrix} \quad D_{14} = \begin{pmatrix} 169 & 86 \\ 73 & 37 \end{pmatrix}$$

$$D_{21} = \begin{pmatrix} 2 & -1 \\ -4 & 3 \end{pmatrix} \quad D_{22} = \begin{pmatrix} -333 & 338 \\ -141 & 147 \end{pmatrix}$$

$$D_{23} = \begin{pmatrix} 338 & 172 \\ 146 & 74 \end{pmatrix}$$

$$D_{31} = \begin{pmatrix} 2 & -1 \\ -4 & 3 \end{pmatrix} \quad D_{32} = \begin{pmatrix} -167 & 169 \\ -69 & 74 \end{pmatrix}$$

$$D_{33} = \begin{pmatrix} 169 & 86 \\ 73 & 37 \end{pmatrix}$$

$$D_{41} = \begin{pmatrix} -166 & 169 \\ -72 & 73 \end{pmatrix} \quad D_{42} = \begin{pmatrix} 169 & 86 \\ 73 & 37 \end{pmatrix}$$

$$D_{51} = \begin{pmatrix} -332 & 338 \\ -144 & 146 \end{pmatrix} \quad D_{52} = \begin{pmatrix} 338 & 172 \\ 146 & 74 \end{pmatrix}$$

$$D_{61} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad (32a)$$

where

$$K_1 = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix} \quad H_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$K_3 = \begin{pmatrix} 1{\cdot}48 & -3{\cdot}44 \\ -1{\cdot}52 & 3{\cdot}56 \end{pmatrix} \quad H_4 = \begin{pmatrix} 0{\cdot}5 & 0 \\ 0 & 0{\cdot}5 \end{pmatrix} \qquad (32b)$$

Because all block elements in the sixth row are null matrices, the block-Routh array terminates prematurely. The greatest common matrix polynomial $C(s)$ is

$$C(s) = C_2 s + C_1 = D_{51} s + D_{52}$$

$$= \begin{pmatrix} -332 & 338 \\ -144 & 146 \end{pmatrix} s + \begin{pmatrix} 338 & 172 \\ 146 & 74 \end{pmatrix} \qquad (32c)$$

It is interesting to notice that the scalar polynomials, $det\,[C(s)] = 200 \times (s^2 + s - 0{\cdot}5)$, is unstable. Using the K_t and H_t in eqn. 32*b* and applying the algorithm in eqn. 13 yields the coprime matrix polynomials as

$$P_1(s) = P_{11} s^2 + P_{12} s + P_{13}$$

$$= \begin{pmatrix} 2{\cdot}18 & -5{\cdot}04 \\ 0{\cdot}72 & -1{\cdot}66 \end{pmatrix} s^2 + \begin{pmatrix} 5{\cdot}74 & 0{\cdot}28 \\ 1{\cdot}24 & 2{\cdot}78 \end{pmatrix} s$$

$$+ \begin{pmatrix} 0{\cdot}5 & 0 \\ 0 & 0{\cdot}5 \end{pmatrix} \qquad (32d)$$

and

$$P_2(s) = P_{21} s + P_{22}$$

$$= \begin{pmatrix} 0{\cdot}74 & -1{\cdot}72 \\ -0{\cdot}76 & 1{\cdot}78 \end{pmatrix} s + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad (32e)$$

Letting $L_1 = I_2$ and substituting K_t and H_t into eqn. 20 gives

$$M_1 = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}$$

$$N_1 = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} 146 & 124 \\ 124 & 106 \end{pmatrix}.$$

Since the $det\,P_{11} (= 0{\cdot}1)$ and $det\,P_{13} (= 0{\cdot}25)$ have the same sign, $N_1 = N_1'$, and M_1 and M_2 are real and positive definite, the system is asymptotically stable. It is interesting to note that the poles of this multivariable system are $s_1 = -0{\cdot}08561$, $s_2 = -0{\cdot}1927$, $s_3 = -6{\cdot}020$ and $s_4 = -251{\cdot}7$. The transmission zeros that are the roots of $det\,[P_2(s)]$ are $s_1 = -0{\cdot}3975$ and $s_2 = -251{\cdot}6025$. The $det\,[D_1(s)] (= det\,P_1(s)\,det\,C(s))$ that consists of unstable eigenvalues is not the characteristic polynomial of this multivariable system.

7 Conclusion

A new block-Routh array with the block-Routh algorithm has been developed to extract the greatest common matrix polynomial of two matrix polynomials that are not coprime, and to construct a block-transformation matrix that transforms a block-state equation from a block-companion form to a block-tridiagonal form. The newly developed block-state equations are the minimal realisations of matrix transfer functions. A driving-point impedance matrix has been synthesised using the structure of multiport *RC* ladder network using the block-state-equation approaches. As a result, the extension of synthesising matrix transfer functions without using integrators is possible. Finally, a stability criterion based on Lyapunov theory has been derived for the test of the stability of a class of matrix transfer functions.

Since we have claimed that the matrix results obtained in this paper are new, it might be interesting to adjust our results from the matrix cases to the scalar cases which are not well known. For single-variable systems, we believe that the simple transformation matrix, which transforms a state equation from a scalar companion form to a scalar tridiagonal form by directly using the elements in the newly developed array, is new. Also, we believe that the derived stability criterion for the transfer function $d_2(s)/d_1(s)$ (in which $d_1(s)$ might have both stable real and stable complex roots and $d_2(s)$ might not be the derivative of $d_1(s)$), is new.

On the other hand, in converting the known scalar results to the matrix cases, the following simple fact should be noticed: the product of two real symmetric, and positive- (or negative-) definite matrices (which are often considered as a natural generalisation of positive (or negative) numbers[26]), often results in a nonsymmetric matrix, and the product of two nonsymmetric matrices may result in a symmetric matrix. For example, the scalar elements $d_{t,1}$ in the first column of the array which is the counterpart of

the block-Routh array are positive (or negative) real values. It is not always true that the block elements $D_{i,1}$ in the first column of the block-Routh array must be symmetric and positive- (or negative-) definite matrices unless the function of interest is a special one. This can be verified from the block elements $D_{i,1}$ in Example 2. The above facts imply that the extension of the known scalar results to the matrix results is not always straightforward.

Finally, it might be interesting to compare the advantages and disadvantages of the present block-Routh array and the matrix-Routh array.[12] We observe that both arrays use similar algorithms but different processes. Therefore, both Routh arrays might have numerically ill-conditioned cases. Also, we observe that the process in the block-Routh array is more complicated than that of the matrix-Routh array. However, the simple and new block-transformation matrix that transforms a block-state equation from a block-companion form to a new block-tridiagonal form of a matrix transfer function can be directly formulated from the block-Routh array but not from the matrix-Routh array. As a result, many applications to circuits and systems have been developed from the new block-tridiagonal matrix that consists of the block quotients obtained from the block-Routh array. Furthermore, the structures of the system matrix, input vector and output vector of the controllable and observable state equations obtained from the block-Routh array are simpler than those of the state equations obtained from the matrix-Routh array. We believe that more applications to circuits and systems can be generated from the present dynamic state equations.

One shortcoming of the method presented is that no precise criterion is offered to ensure the existence of the block-Routh algorithm although some remedial methods have been suggested to overcome the ill-conditioned cases. When an ill-conditioned case occurs, other algorithms, for example, the elementary operation method[30] and the Euler continued fraction method,[27] are more effective in obtaining the greatest common matrix polynomial, determining whether two matrix polynomials are coprime, and checking the stability of a matrix polynomial.

8 Acknowledgments

This work was supported in part by the US Army Research Office, DAAG 29-79-C-0178, and the US Army Missile Research and Development Command, DAAK 40-79-C-0061.

9 References

1 BARNETT, S.: 'Matrices in control theory' (Van Nostrand Reinhold Co., New York, 1971)
2 SHIEH, L.S., and SACHETI, S.: 'A matrix in the Schwarz block form and the stability of matrix polynomials', *Int. J. Control*, 1978, 27, pp. 245–259
3 JURY, E.I.: 'Inners and stability of dynamic systems' (Wiley Interscience, New York, 1974)
4 ANDERSON, B.D.O., JURY, E.I., and MANSOUR, M.: 'Schwarz matrix properties for continuous and discrete time systems', *Int. J. Control*, 1976, 23, pp. 1–16
5 MAROULAS, J., and BARNETT, S.: 'Canonical forms for time-invariant linear control systems', *Int. J. Syst. Sci.*, 1978, 9, pp. 497–514
6 TAKAHASHI, T., HAMADA, N., and TAKAHASHI, S.I.: 'A state-space realisation for transfer functions', *IEEE Trans.*, 1978, CAS-25, pp. 79–88
7 CHEN, C.F., and CHU, H.: 'A matrix for evaluating Schwarz's form', *ibid.*, 1966, AC-11, pp. 303–305
8 BARNETT, S., and STOREY, C.: 'The Lyapunov matix equation and Schwarz's form', *ibid.*, 1967, AC-12, pp. 117–118
9 LOO, S.G.: 'A simplified proof of a transformation matrix relating the companion matrix and the Schwarz matrix', *ibid.*, 1968, AC-13, pp. 309–310
10 POWER, H.M.: 'Canonical form for the matrices of linear discrete-time systems', *Proc. IEE*, 1969, 116, (7), pp. 1245–1252
11 SHIEH, L.S., SHIH, C.D., and YATES, R.E.: 'Some sufficient and some necessary conditions for the stability of multivariable systems', *ASME J. Dyn. Syst. Meas. and Contr.*, 1978, pp. 214–218
12 SHIEH, L.S., and GAUDIANO, F.F.: 'Some properties and application of matrix-continued fraction', *IEEE Trans.*, 1975, CAS-22, pp. 721–728
13 NEWCOMB, R.W.: 'Linear multiport synthesis' (McGraw-Hill, New York, 1966)
14 COOK, M.P., and SHIEH, L.S.: 'A multiport network synthesis using a matrix-continued fraction', *Int. J. Electron.*, 1977, 43, pp. 449–459
15 ANDERSON, B.D.O., and VONGPANITLERD, S.: 'Network analysis and synthesis' (Prentice-Hall, New Jersey, 1973)
16 FRAME, J.S.: 'Matrix functions and applications', *IEEE Spectrum*, 1964, pp. 102–108
17 ROUTH, E.J.: 'A treatise on the stability of a given state of motion' (MacMillan and Co. Ltd., London, 1877)
18 PAPACONSTANTINOU, C.: 'Test for the stability of polynomial matrices', *Proc. IEE*, 1975 122, (3), pp. 312–314
19 ANDERSON, B.D.O., and BITMEAD, R.E.: 'Stability of matrix polynomials', *Int. J. Control*, 1977, 26, pp. 235–247
20 DENMAN, E.D.: 'Matrix polynomials, roots and spectral factors', *Appl. Math. & Comp.*, 1977, 3, pp. 359–368
21 LIAPUNOV, A.M.: 'Stabil v of motion' (Academic Press, Inc., New York, 1966)
22 VAN VALKENBURG, M.E.: 'Network analysis' (Prentice-Hall, New Jersey, 1974)
23 WALL, H.S.: 'Analytic theory of continued fractions' (Chelsea, New York, 1967)
24 GILBERT, E.G.: 'Controllability and observability in multivariable control systems', *SIAM J. Control & Optimiz.*, 1963, 1, pp. 128–151
25 SHIEH, L.S., WEI, Y.J., and NAVARRO, J.M.: 'An algebraic method to determine the common divisor, poles and transmission zeros of matrix transfer functions', *Int. J. Syst. Sci.*, 1978, 9, pp. 949–964
26 BELLMAN, R.: 'Introduction to matrix analysis' (McGraw-Hill, 1970)
27 DESOER, C.A., and VIDYASAGAR, M.: 'Feedback systems: input-output properties' (Academic Press, New York, 1975), pp. 65–82
28 BITMEAD, R.R., and ANDERSON, B.D.O.: 'The matrix Cauchy index: properties and applications', *SIAM J. Appl. Math.*, 1977, 33, pp. 655–672
29 FIELD, A.D., and OWENS, D.H.: 'Canonical form for the reduction of linear scalar systems', *Proc. IEE*, 1978, 125, (4), pp. 337–342
30 ROSENBROCK, H.H.: 'State-space and multivariable theory' (John Wiley and Sons, New York, 1970), pp. 114, 70–220
31 BITMEAD, R.R., KUNG, S.Y., ANDERSON, B.D.O., and KAILATH, T.: 'Greatest common divisors, via generalised Sylvester and Bezout matrices', *IEEE Trans.*, 1978, AC-23, pp. 1043–1047

Corrections to "Analysis and Synthesis of Matrix Transfer Functions Using the New Block-State Equations in Block-Tridiagonal Forms."

L. S. Shieh and A. Tajvari

1) p. 19 (left-hand column)

The phrase in the 9th and 10th lines of section 1 should read: where $\underline{y(t)}$ $\underline{\text{is an m×1 output vector, r(t) is an m×1 input vector}}$, and $x_0(t)$ is an m×1 state vector.

2) p. 20

The block element $(H_2K_3)^{-1}$ in G_1 in Eq. (8c) should read: $(H_2K_1)^{-1}$.

3) p. 23 (right-hand column)

Delete sentences from the 7th line to the 11th line.

4) p. 26 (right-hand column)

The sentence in the 7th line from the bottom of the right-hand column should read: Since P is positive definite $\underline{\text{and}}$ N_i $\underline{\text{are assumed to be imaginary matrices}}$, v is positive definite.

5) p. 27 (left-hand column)

The sentence in the 8th column should read: It is noticed that N_i $\underline{\text{are}}$ imaginary matrices;

6) p. 27 (left-hand column)

The following should be inserted between the 21st and 20th lines from the bottom of the lower left-hand column: $\underline{\text{An additional sufficient condition}}$ is that, if both N_i and M_i are real symmetric matrices and G_2 in Eq. (19) is a real symmetric negative definite matrix, then the system in Eq. (2) is asymptotically stable. This sufficient condition can be verified from the fact that all eigenvalues of a real, symmetric, negative-definite, $\underline{\text{system matrix}}$ G_2 $\underline{\text{in Eq. (19) are negative real.}}$

7) p. 27 (left-hand column)

A phrase is inserted in the 12th and 11th lines from the bottom to read: and the pairs $\{H_{2i}K_{2i-1}\}$ are positive definite $\underline{\text{and the pairs}}$ $\{(K_{2i+1}H_{2i}K_{2i-1})^{-1/2}\}$ $\underline{\text{are imaginary matrices}}$,

8) p. 27 (left-hand column)

A phrase is inserted in the 10th and 9th lines from the bottom to read:

all K_i and H_i are positive definite $\underline{\text{and}}$ G_2 $\underline{\text{in Eq. (19) is a real symmetric negative definite matrix}}$,

9) p. 28 (right-hand column)

A phrase is inserted in the 4th and 3rd lines from the bottom to read: M_1 and M_2 are real symmetric $\underline{\text{and}}$ $G_2<0$ $\underline{\text{in Eq. (19)}}$,

10) p. 30 (right-hand column)

A phrase is inserted in the 9th and 10th lines to read: same sign, $N_1=N_1'$, and M_1 and M_2 are real $\underline{\text{and symmetric and}}$ $G_2<0$ $\underline{\text{in Eq. (19)}}$,

Computer-Aided Methods for Redesigning the Stabilized Pitch Control System of a Semi-Active Terminal Homing Missile

L. S. Shieh[1], M. Datta-Barua[1], R. E. Yates[2] and J. P. Leonard[2].

ABSTRACT

An unstable pitch control system of a terminal homing missile was formerly stabilized using a high order stabilization filter that was realized using active elements. A new dominant-data matching method is presented to redesign the high-order stabilization filter for obtaining reduced-order filters. As a result, the implementation cost is reduced and the reliability increased. An algebraic method is also applied to improve the performance of the redesigned pitch control system. In addition, the proposed dominant-data matching method can be applied to determine a reduced-order model of a high-order system. Unlike most existing model reduction methods, the reduced-order model has the exact assigned frequency-domain specifications of the original system. Computer-aided design methods can also be applied to design general control systems.

---

[1]L. S. Shieh and M. Datta-Barua are with the Department of Electrical Engineering, University of Houston, Houston, Texas 77004.

[2]R. E. Yates and J. P. Leonard are with the Guidance and Control Directorate, U.S. Army Missile Research and Development Command, Redstone Arsenal, Alabama 35809.

A Geometric Series Approach for Approximation of Transition Matrices in Quadratic Synthesis

Leang S. Shieh,[1] Willon B. Wai,[1] R. E. Yates[2]

Abstract

A geometric-series approach is used to approximate the exponentials of Hamiltonian matrices for quadratic synthesis problems. The approximants of the discretized transition matrices are then used to construct piecewise-constant gains and piecewise-time varying gains for approximating a time-varying optimal gain and a time-varying Kalman gain. The proposed method is more accurate and computationally faster than those existing methods which use the Walsh function approach and the block-pulse function approach.

[1] Department of Electrical Engineering, University of Houston, Houston, Texas 77004.

[2] Guidance and Control Directorate, U. S. Army Missile Command, Redstone Arsenal, Alabama 35809.